Demographic and Health Survey

Sampling and Household Listing Manual

ICF International
Calverton, Maryland USA

September 2012

MEASURE DHS is a five-year project to assist institutions in collecting and analyzing data needed to plan, monitor, and evaluate population, health, and nutrition programs. MEASURE DHS is funded by the U.S. Agency for International Development (USAID). The project is implemented by ICF International in Calverton, Maryland, in partnership with the Johns Hopkins Bloomberg School of Public Health/Center for Communication Programs, the Program for Appropriate Technology in Health (PATH), Futures Institute, Camris International, and Blue Raster.

The main objectives of the MEASURE DHS program are to: 1) provide improved information through appropriate data collection, analysis, and evaluation; 2) improve coordination and partnerships in data collection at the international and country levels; 3) increase host-country institutionalization of data collection capacity; 4) improve data collection and analysis tools and methodologies; and 5) improve the dissemination and utilization of data.

For information about the Demographic and Health Surveys (DHS) program, write to DHS, ICF International, 11785 Beltsville Drive, Suite 300, Calverton, MD 20705, U.S.A. (Telephone: 301-572-0200; fax: 301-572-0999; e-mail: info@measuredhs.com; Internet: http://www.measuredhs.com).

Recommended citation:
ICF International. 2012. *Demographic and Health Survey Sampling and Household Listing Manual*. MEASURE DHS, Calverton, Maryland, U.S.A.: ICF International

TABLE OF CONTENTS

TABLES AND FIGURES ..... **vii**

1 DEMOGRAPHIC AND HEALTH SURVEYS SAMPLING POLICY **1**

1.1 General principles 1

1.1.1 Existing sampling frame 1

1.1.2 Full coverage 1

1.1.3 Probability sampling 2

1.1.4 Suitable sample size 2

1.1.5 Simple design 2

1.1.6 Household listing and pre-selection of households 2

1.1.7 Good sample documentation 2

1.1.8 Confidentiality 3

1.1.9 Exactness of survey implementation 3

1.2 Survey objectives and target population 3

1.3 Survey domain 4

1.4 Sampling frame 4

1.4.1 Conventional sampling frame 5

1.4.2 Alternative sampling frames 5

1.4.3 Evaluation of the sampling frame 6

1.5 Stratification 6

1.6 Sample size 7

1.6.1 Sample size and sampling errors 7

1.6.2 Sample size determination 10

1.7 Sample allocation 12

1.8 Two-stage cluster sampling procedure 15

1.9 Sample “take” per cluster 16

1.9.1 Optimum sample take 16

1.9.2 Variable sample take for self-weighting 17

1.10 Household listing 19

1.11 Household selection in the central office 20

1.12 Household interviews 21

1.13 Sampling weight calculation 22

1.13.1 Why we need to weight the survey data 22

1.13.2 Design weights and sampling weights 22

1.13.3 How to calculate the design weights 23

1.13.4 Correction of unit non-response and calculation of sampling weights....24
1.13.5 Normalization of sampling weights....26
1.13.6 Standard weights for HIV testing....27
1.13.7 De-normalization of standard weights for pooled data....28
1.14 Calibration of sampling weights in case of bias....29
1.15 Data quality and sampling error reporting....30
1.16 Sample documentation....31
1.17 Confidentiality....31
2 HOUSEHOLD LISTING OPERATION....32
2.1 Introduction....32
2.2 Definition of terms....32
2.3 Responsibilities of the listing staff....33
2.4 Locating the cluster....34
2.5 Preparing location and sketch maps....35
2.6 Collecting a GPS waypoint for each cluster....36
2.7 Listing of households....37
2.8 Segmentation of large clusters....38
2.9 Quality control....39
2.10 Prepare the household listing forms for household selection....39
Appendix 2.1 Example listing forms....41
Appendix 2.2 Symbols for mapping and listing....46
Appendix 2.3 Examples of completed mapping and listing forms....48
3 SELECTED SAMPLING TECHNIQUES....52
3.1 Simple random sampling....52
3.2 Equal probability systematic sampling....53
3.2.1 Sampling theory....53
3.2.2 Excel templates for systematic sampling....55
3.3 Probability proportional to size sampling....64
3.3.1 Sampling theory....64
3.3.2 Operational description and examples....65
3.4 Complex sampling procedures....70
4 SURVEY ERRORS....73
4.1 Errors of coverage and non-response....73
4.1.1 Coverage errors....73
4.1.2 Deliberate restrictions of coverage....74
4.1.3 Non-response....74

4.1.4 Response rates 76
4.2 Sampling errors 78
5 SAMPLE DOCUMENTATION 80
5.1 Introduction 80
5.2 Sample design document 80
5.2.1 Introduction 80
5.2.2 Sampling frame 81
5.2.3 Structure of the sample and the sampling procedure 82
5.2.4 Selection probability and sampling weight 84
5.3 Sample file 85
5.4 Results of Survey implementation 88
5.5 Sampling errors 90
5.6 Sampling parameters in DHS data files 91
Glossary of terms 93
References 97

TABLES AND FIGURES

Table 1.1 Sample size determination for estimating current use of a modern contraceptive method among currently married women....10

Table 1.2 Sample size determination for estimating the prevalence of full vaccination coverage among children aged 12-23 months....11

Table 1.3 Sample allocation: Proportional allocation....14

Table 1.4 Sample allocation: Power allocation....14

Table 1.5 Optimal sample take for currently married women 15-49 currently using any contraceptive method based on intracluster correlation ρ and survey cost ratio c_1/c_2 from past surveys....17

Table 5.1 Distribution of EAs and average size of EA by region and by type of residence....82

Table 5.2 Distribution of households by region and by type of residence....82

Table 5.3 Sample allocation of clusters and households by region and by type of residence....84

Table 5.4 Expected number of interviews by region and by type of residence....84

Table 5.5 An example sample file....87

Table 5.6 Example table for the results of survey implementation....88

Table 5.7 Example appendix table for the results of the women's survey implementation....89

Table 5.8 Example appendix table for the results of the men's survey implementation....90

Table 5.9 Example table for sampling errors....91

Figure 3.1 Simple household selection with a sub-sample....57

Figure 3.2 Selection of runs with a sub-sample....58

Figure 3.3 Simple self-weighting selection without sample size control....59

Figure 3.4 Self-weighting selection with runs and without sample size control....60

Figure 3.5 Self-weighting selection with sample size control....61

Figure 3.6 Self-weighting selection with runs and with sample size control....62

Figure 3.7 Manual household selection in the field....63

Figure 3.8 Part of an Excel template for stratified sampling....68

Figure 3.9 Part of an example for a province crossed urban-rural stratified PPS sampling....69

Figure 3.10 Part of an example sample file from a stratified PPS sampling....70

1 DEMOGRAPHIC AND HEALTH SURVEYS SAMPLING POLICY

1.1 General principles

Scientific sample surveys are cost-efficient and reliable ways to collect population-level information such as social, demographic and health data. The MEASURE DHS project is a worldwide project implemented across various countries and at multiple points in time within a country. In order to achieve **comparability, consistency** and the **best quality** in survey results, sampling activities in the Demographic and Health Surveys (DHS) should be guided by a number of general principles. This manual presents general guidelines on sampling for DHS surveys, although modifications may be required for country-specific situations. The key principles of DHS sampling include:

- Use of an existing sampling frame
- Full coverage of the target population
- Probability sampling
- Using a suitable sample size
- Using the most simple design possible
- Conducting a household listing and pre-selection of households
- Providing good sample documentation
- Maintaining confidentiality of individual's information
- Implementing the sample exactly as designed

1.1.1 Existing sampling frame

A probability sample can only be drawn from an existing sampling frame which is a complete list of statistical units covering the target population. Since the construction of a new sampling frame is likely to be too expensive, DHS surveys should use an adequate pre-existing sampling frame which is officially recognized. This is possible for most of the countries where there has been a population census in recent years. Census frames are generally the best available sampling frame in terms of coverage, cartographic materials and organization. However, an evaluation of the quality and the accessibility of the frame should be considered during the development of the survey design, and a detailed study of the sampling frame is necessary before drawing the sample. In the absence of a census frame, a DHS survey can use an alternative sampling frame, such as a complete list of villages or communities in the country with all necessary identification information including a measure of population size (e.g. number of households), or a master sample which is large enough to support the DHS design.

1.1.2 Full coverage

A DHS survey should cover 100 percent of the target population in the country. The target population for the DHS survey is all women age 15-49 and children under five years of age living in residential households. Most surveys also include all men age 15-59[1]. The target population may vary from country to country or from survey to survey, but the general sampling principles are the same. In some cases, exclusion of some areas may be necessary because of extreme inaccessibility, violence or instability, but these issues need to be considered at the very beginning of the survey, before the sample is drawn.

[1] The age range varies from survey to survey and may be 15-49, 15-54, 15-59 or 15-64.

1.1.3 Probability sampling

A scientific probability sampling methodology must be used in DHS surveys. A probability sample is defined as one in which the units are selected randomly with known and nonzero probabilities. This is the only way to obtain unbiased estimation and to be able to evaluate the sampling errors. The term probability sampling excludes purposive sampling, quota sampling, and other uncontrolled non-probability methods because they cannot provide evaluation of precision and/or confidence of survey findings.

1.1.4 Suitable sample size

Sample size is a key parameter for DHS surveys because it is directly related to survey budget, data quality and survey precision. Theoretically, the larger the sample size, the better the survey precision, but this is not always true in practice. Survey budget is not the only important factor in determining the sample size. Desired precision, the number of domains, capability of the implementing organization, data quality concerns and cost effectiveness are essential constraints in determining the total sample size. Thus a suitable sample size is also a key parameter to guarantee data quality.

1.1.5 Simple design

In large-scale surveys, non-sampling errors (coverage errors, errors committed in survey implementation and data processing, etc.) are usually the most important sources of error and are expensive to control and difficult to evaluate quantitatively. It is therefore important to minimize them in survey implementation. In order to facilitate accurate implementation of the survey, the sampling design for DHS should be as simple and straightforward as possible. Macro's experience from 25 years of DHS surveys shows that a two-stage household-based sample design is relatively easy to implement and that quality can be maintained.

1.1.6 Household listing and pre-selection of households

The DHS standard procedure recommends that households be pre-selected in the central office prior to the start of fieldwork rather than by teams in the field who may have pressures to bias the selection. The interviewers are asked to interview only the pre-selected households. In order to prevent bias, no changes or replacements are allowed in the field. To perform pre-selection of households, a complete list of all residential households in each of the selected sample clusters is necessary. This list is usually obtained from a household listing operation conducted before the main survey.

In some surveys, the household listing operation may be combined with the main survey to form a single field operation, and households can be selected in the field from a complete listing. Combining the household listing and survey data collection in one field operation is less expensive; however, it provides incentive to leave households off the household list to reduce workload, thus reducing the representativeness of the survey results. Close supervision is needed during the field work to prevent this problem. Separate listing and data collection operations are thus required for this reason. Interviewers selecting households in the field without a complete listing is not acceptable for DHS surveys.

1.1.7 Good sample documentation

DHS surveys are usually year-long projects conducted by different people specialized in different aspects of survey implementation, so good sample documentation is necessary to guarantee the exact implementation of the project. The sample documentation should include a sample design document and the list of primary sampling units. The sample design document should explain in detail

the methodology, the sampling procedure, the sample size, the sample allocation, the survey domains and the stratification. This should also form the basis for an appendix to the DHS final report describing the sample design. The sample list should include all identification information for all of the selected sample points, along with their probability of selection.

1.1.8 Confidentiality

Confidentiality is a major concern in DHS, especially when human bio-markers are collected such as blood samples for HIV testing. The DHS surveys are anonymous surveys which do not allow any potential identification of any single household or individual in the data file. Confidentiality is also a key factor affecting the response rate to sensitive questions regarding sexual activity and partners.

In particular, in surveys that include HIV testing DHS policy requires that PSU and household codes are scrambled in the final data to further anonymize the data and the original sample list is destroyed.

1.1.9 Exactness of survey implementation

Exactness of sample implementation is the last element in achieving good sampling precision. No matter how carefully a survey is designed and how complete the materials for conducting sampling activities are, if the implementation of the sampling activities by sampling staff (office staff responsible for selecting sample units, field workers responsible for the mapping and household listing and interviewers responsible for data collection) is not preformed exactly as designed, serious bias and misleading results may occur.

In the sections that follow, DHS policies related to sample design and implementation are described.

1.2 Survey objectives and target population

The main objective of DHS surveys is to collect up-to-date information on basic demographic and health indicators, including housing characteristics, fertility, childhood mortality, contraceptive knowledge and use, maternal and child health, nutritional status of mothers and children, knowledge, attitudes and behavior toward HIV/AIDS and other sexually transmitted infections (STI), women's status. The target population for DHS is defined as all women of reproductive age (15-49 years old) and their young children under five years of age living in ordinary residential households. However, in some countries, the coverage may be restricted to ever-married women.

The main indicator topics include:

- Total fertility and age specific fertility rates
- Age at first sex, first birth, and first marriage
- Knowledge and use of contraception
- Unmet need for family planning
- Birth spacing
- Antenatal care
- Place of delivery
- Assistance from skilled personnel during delivery
- Knowledge of HIV/AIDS and other STIs
- Higher-risk sexual behavior
- Condom use
- Childhood vaccination coverage
- Treatment of diarrhea, fever, and cough

- Infant and under-five mortality rates
- Nutritional status

Since the target population can be easily found in residential households, DHS is a household-based survey.

1.3 Survey domain

In DHS surveys, an important objective is to compare the survey results for different characteristics such as urban and rural residence, different administrative or geographic regions, or different educational levels of respondents. A *survey domain* or *study domain* is a sub-population for which separate estimation of the main indicators is required. There are two kinds of survey domains: *design domains* and *analysis domains*. A design domain consists of a sub-population which can be identified in the sampling frame and therefore can be handled independently in the sample size and sampling procedures, usually consisting of geographic areas or administrative units. For example, urban and rural differences are very frequently requested; therefore, urban and rural areas are usually separate design domains for Demographic and Health Surveys. An analysis domain is a sub-population which cannot be identified in the sampling frame, such as domains specified by individual characteristics. These may include women with secondary or higher education, pregnant women, children 12-23 months, and children having diarrhea in the two weeks preceding the survey.

In order for survey estimates to be reliable at the domain level, it is necessary to ensure that the number of cases in each survey domain is sufficient, especially when desired levels of precision are required for particular domains. For a design domain, adequate sample size is achieved by allocating the target population at the survey design stage into the requested design domains, and then calculating the sample size for the specific design domains by taking the precision required into account. On the other hand, for an analysis domain, it is difficult to guarantee a specified precision because it is difficult to control the sample size at the design stage. However, if prior estimates of the average number of target individuals per household are available, then it is possible to control the precision for an analysis domain. For example, if survey estimates are required for the nutritional status of children under age 5 is required and estimates of the number of children under age 5 per household are available, it is then possible to calculate a sample size to give a certain level of precision.

DHS reports also produce some indicators for *second level domains* such as vaccination coverage of children age 12-23 months within a region, where region is the first level domain, and children 12-23 months is the second level domain. Caution must be paid to the precision required for a second level domain because the second level domain usually includes a very small sub-population.

If domain-level estimates are required, it is better to avoid a large number of domains because otherwise a very large sample size will be needed. The number of domains and the desired level of precision for each must be taken into account in the budget calculation and assessment of the implementation capabilities of the implementing organization. The total sample size needed is the sum of sample sizes needed in all exclusive (first level) domains.

1.4 Sampling frame

A *sampling frame* is a complete list of all *sampling units* that entirely covers the target population. The existence of a sampling frame allows a probability selection of sampling units. For a multi-stage survey, a sampling frame should exist for each stage of selection. The sampling unit for the first stage of selection is called the *Primary Sampling Unit* (PSU); the sampling unit for the second stage of selection is called the *Secondary Sampling Unit* (SSU), and so on. In most cases, DHS surveys

are two-stage surveys. Note that each stage of sample selection will involve sampling errors, so it is better to avoid more than two stages if additional stages of selection are not necessary.

The availability of a suitable sampling frame is a major determinant of the feasibility of conducting a DHS survey. This issue should be addressed in the earliest stages of planning for a survey. A sampling frame for a DHS survey could be an existing sampling frame, an existing master sample, or a sample of a previously executed survey of sufficiently large sample size, which allows for the selection of subsamples of desired size for the DHS survey.

1.4.1 Conventional sampling frame

The best frame is the list of *Enumeration Areas* (EAs) from a recently completed population census. An EA is usually a geographic area which groups a number of households together for convenient counting purposes for the census. A complete list of EAs which covers the survey area entirely is the most ideal frame for DHS surveys.

In most cases, a list of EAs from a recent census is available. This list should be thoroughly evaluated before it is used. The sampling frame used for DHS should be as up-to-date as possible. It should cover the whole survey area, without omission or overlap. Basic cartographic materials should exist for each area unit or at least for groups of units with clearly defined boundaries. Each area unit should have a unique identification code or a series of codes that, when combined, can serve as a unique identification code. Each unit should have at least one measure of size estimate (population and/or number of households). If other characteristics of the area units (e.g., socioeconomic level) exist, they should be evaluated and retained as they may be used for stratification.

A pre-existing master sample (which is a random sample from the census frame) can be accepted only where there is confidence in the master sample design, including detailed sampling design parameters such as sampling method, stratification, and inclusion probability for the selected primary sampling units. The task for the DHS survey is then to design a sub-sampling procedure, which produces a sample in line with DHS requirements. This will not always be possible. However, the larger the master sample is in relation to the desired DHS sub-sample, the more flexibility there will be for developing a sub-sampling design. A key question with a pre-existing sample is whether the listing of dwellings/households is still current or whether it needs to be updated. If updating is required, use of a pre-existing sample may not be economical. The potential advantages of using a pre-existing sample are: 1) economy, and 2) increased analytic power through comparative analysis of two or more surveys. The disadvantages are: 1) the problem of adapting the sample to DHS requirements, and 2) the problem of repeated interviews with the same household or person in different surveys, resulting in respondent fatigue or contamination. One way to avoid this last problem is to keep just the primary sampling units from the pre-existing sample and reselect the households for the DHS survey.

1.4.2 Alternative sampling frames

When neither a census frame nor a master sample is available then alternative frames should be considered. Examples of such frames are:

- A list of electoral zones with estimated number of qualified voters for each zone
- A gridded high resolution satellite map with estimated number of structures for each grid
- A list of administrative units such as villages with estimated population for each unit

A main concern when using alternative frames are coverage problems, that is, does the frame completely cover the target population? Usually checking the quality of an alternative frame is more difficult because of a lack of information either from the frame itself or from administrative sources. Another problem is the size of the primary sampling unit. Since the alternative frame is not specifically

created for a population census or household based survey, the size of the PSUs of such frames may be too large or too small for a DHS survey. A third problem is identifying the boundaries of the sampling units due to the lack of cartographic materials.

In the first two examples of alternative sampling frames, the standard DHS two-stage sampling procedure can be applied by treating the electoral zones or the grids of satellite map as the PSUs. In the third case, when a list of administrative units larger than villages (e.g. sub-districts, wards or communes) is available, for example, a complete list of all communes in a country may be easier to get than a complete list of villages, then it is necessary to use a selection procedure that includes more than two stages. In the first stage, select a number of communes; in each of the selected communes, construct a complete list of all villages residing in the commune; select one village per commune as a DHS cluster, then proceed with the subsequent household listing and selection as in a standard DHS. This procedure works best when the number of communes is large and the commune size is small. A list of administrative units that are small in number but large in size is not suitable for a DHS sampling frame because this situation will result in large sampling errors, as explained later in Section 1.9.

1.4.3 Evaluation of the sampling frame

No matter what kind of sampling frame will be used, it is always necessary to check the quality of the frame before selecting the sample. Following are several things that need to be checked when using a conventional sampling frame:

- Coverage
- Distribution
- Identification and coding
- Measure of size
- Consistency

There are several easy but useful ways to check the quality of a sampling frame. For example, for a census frame, check the total population of the sampling frame and the population distribution among urban and rural areas and among different regions/administrative units obtained from the frame with that from the census report. Any important differences may indicate that there may be coverage problems. If the frame provides information on population and households for each EA, then the average number of household members can be calculated, and a check for extreme values can help to find incorrect measures of size of the PSUs. If information on population by sex is available for each EA, then a sex ratio can be calculated for each EA, and a check for extreme values can help to identify non-residential EAs. If the EAs are associated with an identification (ID) code, then check the ID codes to identify miscoded or misplaced EAs. A sampling frame with full coverage and of good quality is the first element for a DHS survey; therefore, efforts should be made to guarantee a good start for the project.

For a nationally representative survey, geographic coverage of the survey should include the entire national territory unless there are strong reasons for excluding certain areas. If areas must be excluded, they should constitute a coherent domain. A survey from which a number of scattered zones have been excluded is difficult to interpret and to use.

1.5 Stratification

Stratification is the process by which the survey population is divided into subgroups or strata that are as homogeneous as possible using certain criteria. *Explicit stratification* is the actual sorting and separating of the units into specified strata. Within each stratum, the sample is designed and selected independently. It is also possible to systematically sample units from an ordered list (with a

fixed sampling interval between selected units) to achieve the effect of stratification. For example, in DHS survey, it is not unusual for the PSUs within the explicit strata to be sorted geographically. This is called *implicit stratification*.

The principal objective of stratification is to reduce sampling errors. In a stratified sample, the sampling errors depend on the population variance existing within the strata but not between the strata. For this reason, it pays to create strata with low internal variability (or high homogeneity). Another major reason for stratification is that, where marked differences exist between subgroups of the population (e.g., urban vs. rural areas), stratification allows for a flexible sample design that can be different for each subgroup.

Stratification should be introduced only at the first stage of sampling. At the dwelling/household selection stage, systematic sampling is used for convenience; however, no attempt should be made to reorder the dwelling/household list before selection in the hope of increasing the implicit stratification effect. Such efforts generally have a negligible effect.

Stratification can be single-level or multi-level. In single-level stratification, the population is divided into strata according to certain criteria. In multi-level stratification, the population is divided into first-level strata according to certain criteria, and then the first-level strata are subdivided into second-level strata, and so on. A typical two-level stratification involves first stratifying the population by region at the first level and then by urban-rural within each region. A DHS survey usually employs multi-level stratification.

Strata should not be confused with survey domains. A survey domain is a population subgroup for which separate survey estimates are desired (e.g., urban areas/rural areas). A stratum is a subgroup of homogeneous units (e.g., subdivisions of an administrative region) in which the sample may be designed differently and is selected separately. Survey domains and strata can be the same but they need not be. For example, survey domains could be the first-level stratum in a multi-level stratification. On the other hand, a survey domain could consist of one or several lower-level strata.

DHS surveys typically use explicit stratification by separating urban and rural residence within each region. Where data are available, explicit stratification could also be done on the basis of socio-economic zones or more directly relevant characteristics such as the level of female literacy or the presence of health facilities in the areas. These kinds of information could be obtained from administrative sources. Within each explicit stratum, the units can then be ordered according to location, thus providing further implicit geographic stratification.

1.6 Sample size

1.6.1 Sample size and sampling errors

The estimates from a sample survey are affected by two types of errors: *sampling errors* and *non-sampling errors*. Sampling errors are the representative errors due to sampling of a small number of eligible units from the target population instead of including every eligible unit in the survey. Sampling errors are related to the sample size and the variability among the sampling units. Sampling errors can be statistically evaluated after the survey. Non-sampling errors result from problems during data collection and data processing, such as failure to locate and interview the correct household, misunderstanding of the questions on the part of either the interviewer or the respondent, and data entry errors. Non-sampling errors are related to the capacity of the implementing organization, and experience shows that (1) non-sampling errors are always the **most important source of error** in a survey, and (2) it is difficult to evaluate the magnitude of non-sampling errors once a survey is complete. Theoretically, with the same survey methodology and under the same survey conditions, the larger the sample size, the better the survey precision. However, this relationship does not always hold

true in practice, because non-sampling errors tend to increase with survey scale and sample size. The challenge in deciding on the sample size for a survey is to balance the demands of analysis and precision with the capacity of the implementing organization and the constraints of funding.

A common measure of precision for estimating an indicator is its *relative standard error* (RSE) which is defined as its *standard error* (SE) divided by the estimated value of the indicator. The standard error of an estimator is the representative error due to sampling. The relative standard error describes the amount of sampling error relative to the indicator level and is independent of the scale of the indicator to be estimated; therefore, a unique RSE can be applied to a reference indicator for all domains. If a unique RSE is desired for all domains, the domain sample size depends on the variability and the size of the domain. The total sample size is the sum of the sample sizes over all domains for which desired precision are required. The following are some concepts related to sample size calculation.

1. The standard error of an estimator when estimating a proportion with a *simple random sampling without replacement*[2] is given by:

$$SE = SQRT\left(\frac{1\text{-f}}{n} \times \frac{N}{N-1} \times P(1-P)\right)$$

where n is the sample size (number of completed interviews),
P is the proportion,
N is the target population size, and
f=n/N is the sampling fraction.

When N is large and n is relatively small, the above quantity can be approximated by:

$$SE \simeq SQRT\left(\frac{P(1-P)}{n}\right)$$

Therefore the RSE of the estimator is given by:

$$RSE(P) \simeq SQRT\left(\frac{P(1-P)}{n}\right) / P = SQRT\left(\frac{1/P-1}{n}\right)$$

2. For a required precision with a relative standard error α, the net sample size (number of completed interviews) needed for a simple random sampling is given by:

$$n = \frac{(1/P-1)}{\alpha^2}$$

3. Since a simple random sampling is not feasible for a DHS, the sample size for a complex survey with clustering such as the DHS can be calculated by inflating the above calculated sample size by using a design effect (*Deft)*. *Deft* is a measure of efficiency of cluster sampling compared to a direct simple random sampling of individuals, defined as the ratio between the standard error using the given sample design and the standard error that would result if a

[2] A simple random sample would be a random selection of individuals or households directly from the target population. This is not feasible for DHS surveys because a list of all eligible individuals or households is not available.

simple random sample had been used. A *Deft* value of 1.0 indicates that the sample design is as efficient as a simple random sample, while a value greater than 1.0 indicates the increase in the sampling error due to the use of a more complex and less statistically efficient design. The net sample size needed for a cluster sampling with same relative standard error is given by:

$$n = \text{Deft}^2 \times \frac{(1/P - 1)}{\alpha^2}$$

4. The formula for calculating the final sample size in terms of the number of households while taking non-response into account (the formula used in the templates for sample size calculation as shown in Table 1.1) is given by:

$$n = \text{Deft}^2 \times \frac{(1/P - 1)}{\alpha^2} / (R_i \times R_h \times d)$$

where n is the sample size in households;
Deft is the design effect (a default value of 1.5 is used for Deft if not specified);
P is the estimated proportion;
α is the desired relative standard error;
R_i is the individual response rate;
R_h is the *household gross response rate*; and
d is the number of eligible individuals per household.

The *household gross response rate* is the number of households interviewed over the number selected. DHS reports typically report the net household response rate which is the number of households interviewed over the number valid households found in the field (i.e. excluding vacant and destroyed dwellings.)

5. If the target population is small (such as in a sub-national survey), a finite population correction of the above calculated sample size should be applied. The final sample size n is calculated by

$$n = \frac{n_0}{1 + n_0/N}$$

where n_0 is the initial sample size calculated in point number 4, and N is the target population size.

6. The relationship between the RSE and the sample size shows that, if one reduces a desired RSE to half, then the sample size needed will increase 4 times. For example, the sample size for a RSE of 5% is 4 times larger than the sample size for a RSE of 10% (see Tables 1.1 and 1.2 in the next section). This means that it is very expensive to reduce the RSE by increasing the sample size. Therefore, when designing the sample size, the efficiency of the design must be considered, that is, the balance between the gain in precision and the increase in sample size (or survey cost).

7. The width of the confidence interval is determined by the RSE. With a confidence level of 95%, 2*P*RSE is the half-length of the confidence interval for P. For example, for RSE=0.10 and P=0.20, the half-length of the confidence interval is 0.04, which means the confidence interval for P is (0.16, 0.24). (DHS reports +/-2*SE instead of +/-1.96*SE as 95% confidence interval for conservative purposes).

1.6.2 Sample size determination

The total sample size for a DHS survey with a number of survey domains (design domain) is the sum of the sample sizes over all domains. An appropriate sample size for a survey domain is the minimum number of persons (e.g., women age 15-49, currently married women 15-49, children under age five) that achieves the desired survey precision for core indicators at the domain level. If funding is tight and fixed, the sample size is the maximum number of persons that the funding can cover. Precision at the national level is usually not a problem. In almost all cases, sample size is decided to guarantee precision at domain level with appropriate allocation of the sample. So apart from survey costs, the total sample size depends on the desired precision at domain level and the number of domains. If a reasonable precision is required at domain level, experience from the MEASURE DHS program shows that a minimum number of 800 completed interviews with women is necessary for some of the woman-based indicators for high fertility countries (e.g. total fertility rate, contraceptive prevalence rate, childhood mortality rates); for low fertility countries, the minimum domain sample size can reach 1,000 completed interviews or more. Table 1.1 below illustrates the calculation of sample size for a domain according to different levels of desired RSE for estimating the indicator “the proportion of currently married women who are current users of a modern contraceptive method”.

Table 1.1 Sample size determination for estimating current use of a modern contraceptive method among currently married women

Estimated proportion p		0.20	Total target population		
Estimated design effect (Deft)		1.40	# of target individuals/HH		1.05
Individual response rate		0.96	HH gross response rate		0.92
Desired RSE	**Net Sample size individual**	**Sample size Household**	**Expected SE**	**95% confidence limits**	
				Lower	**Upper**
0.20	196	212	0.040	0.120	0.280
0.19	217	234	0.038	0.124	0.276
0.18	242	261	0.036	0.128	0.272
0.17	271	293	0.034	0.132	0.268
0.16	306	330	0.032	0.136	0.264
0.15	348	376	0.030	0.140	0.260
0.14	400	432	0.028	0.144	0.256
0.13	464	501	0.026	0.148	0.252
0.12	544	587	0.024	0.152	0.248
0.11	648	699	0.022	0.156	0.244
0.10	784	846	0.020	0.160	0.240
0.05	3136	3382	0.010	0.180	0.220

Note: The confidence limits are calculated as P±2*SE.

Assuming the domain size is large enough such that the finite population correction is negligible, Table 1.1 gives the required gross sample size in terms of number of households with estimated parameters from a DHS survey. The target population is currently married women age 15-49; the estimated parameters are:

- the proportion of currently married women who are current users of any modern contraceptive method,
- the design effect (Deft),
- the number of target individuals (number of currently married women 15-49) per household,
- the individual and the household response rates.

For example, with an estimated prevalence of 20%, if we require a RSE of 10%, we should select 846 households in this particular domain. With a gross household response rate (the number of households completed over the total number selected) of 92% and an individual response rate of 96%, we expect to obtain 784 completed interviews of currently married women age 15-49.

The estimated quantities at the top of the table used as input to the calculation can usually be obtained from previous surveys or from administrative records. The total sample size for a survey with several domains is the sum of the sample sizes obtained in the above table for each domain. If the same precision required and the same indicator level apply to all domains, then the total sample size is the sample size calculated for one domain multiplied by the number of domains. With this example, the total sample size for a survey having six domains with approximately the same level of modern contraceptive use among currently married women and the same precision request for each domain would be 5076 households. The “Sample size determination” template located in the Appendix can be used to determine required sample sizes.

Table 1.2 Sample size determination for estimating the prevalence of full vaccination coverage among children aged 12-23 months

Estimated proportion p		0.29	Total target population		
Estimated design effect (Deft)		1.22	# of target individuals/HH		0.11
Individual response rate		0.96	HH gross response rate		0.92
Desired RSE	**Net Sample size individual**	**Sample size household**	**Expected SE**	**95% confidence limits**	
				Lower	**Upper**
0.20	91	937	0.058	0.174	0.406
0.19	101	1040	0.055	0.180	0.400
0.18	112	1153	0.052	0.185	0.395
0.17	126	1297	0.049	0.191	0.389
0.16	142	1462	0.046	0.197	0.383
0.15	162	1668	0.043	0.203	0.377
0.14	186	1915	0.041	0.209	0.371
0.13	216	2224	0.038	0.215	0.365
0.12	253	2605	0.035	0.220	0.360
0.11	301	3099	0.032	0.226	0.354
0.10	364	3747	0.029	0.232	0.348
0.05	1458	15008	0.014	0.261	0.319

Note: The default value of Deft is set to be 1.5. Specify if different.

The confidence limits are calculated as P±2*SE.

If response rate is not provided, the sample size calculated is net sample size.

Table 1.2 shows a similar example for the indicator “proportion of children aged 12-23 months who are fully immunized”. In this case, the target population is children aged 12-23 months. The estimated number of target individuals per household is much smaller than the number of currently married women per household given in Table 1.1. So for the same sample size calculated in Table 1.1, we can only get a RSE of above 20% at domain level. With a RSE of 10%, we need to select 3746 households in this particular domain which seems unrealistic if we have several domains for the survey.

This example shows that for a multi-indicator survey, the sample size required can be very different from indicator to indicator. So the choice of the reference indicator upon which the sample size is calculated is an important issue. The reference indicator which is used for sample size determination should have demographic importance, moderate value and moderate population coverage, i.e. apply to a sizable proportion of the population. With the same sample size calculated in Table 1.1 for a survey having six domains, the RSE for the whole sample for estimating full immunization among children 12-23 months is between 8% and 9%.

The domain sample sizes often need to be balanced between domains due to budget constraints. In practice it is often the case that the total sample size is fixed according to funding available and implementation capacity, and then the sample is allocated to each domain and to each stratum within the domain. In the case of very tight budget constraints, we may equally allocate the total sample to the domains. In some cases, we may want to oversample a specific domain to conduct some in-depth analysis for a certain rare phenomenon. The method (and the tables) presented in the following section may be used to allocate the sample at the domain level because the domains are usually first-level strata. Regardless of the method used for allocation, the calculation of domain sample size can give us an idea about the precision we may achieve in each domain with a given sample size.

1.7 Sample allocation

In cases where the total sample size or domain sample size has been fixed, we need to appropriately allocate the sample to different domains (or different strata within a domain). This allocation is aimed at strengthening the sampling efficiency at the national level or domain level and reducing sampling errors. Assuming a constant cost across domains/strata, the optimum allocation of the sample depends on the size of the domain/stratum N_h and the variability of the indicator to be estimated S_{xh}'''

$$n_h \propto N_h' S_{xh}'''$$

For a given total sample size n the optimum allocation for variable x is given by:

$$n_h = n'' \frac{N_h' S_{xh}}{\sum_{h=1}^{H} N_h' S_{xh}}$$

The optimum allocation is only optimal for the indicator on which the allocation is based; that allocation may not be appropriate for other indicators. For a multipurpose survey, if the domains/strata are not too different in size, a safe allocation that is good for all indicators is a proportional allocation, with sample size proportional to the domain/stratum size.

$$n_h = n'' \frac{N_h'}{\sum_{h=1}^{H} N_h'} = n \frac{N_h}{N}$$

This allocation introduces a constant sampling fraction across domain/strata with:

$$f_h = \frac{n_h}{N_h} = \frac{n}{N}$$

Because DHS surveys are multipurpose surveys, a proportional allocation of sample is recommended if the domains/strata are not too different in size. However, if the domains/strata sizes are very different, the smaller domains/strata may receive a very small sample size.

If a desired precision is required at domain/stratum level, by assuming equal relative variations across strata, a power allocation (Bankier, 1988) with an appropriate power value α ($0 \le \alpha \le 1$) may be used to guarantee sufficient sample size in small domains/strata.

$$n_h = n'' \frac{M_h^{\alpha\prime}}{\sum_{h=1}^{H} M_h^{\alpha\prime}}$$

A power allocation is an allocation proportional to the power of a size measure *M*. A power value of 1 gives proportional allocation; a power value of 0 gives equal size allocation; a power value between 0 and 1 gives an allocation between proportional allocation and equal size allocation. Proportional allocation is good for national level indicators, but may not meet the precision request at domain level; while an equal size allocation is good for comparison across domains, but may affect the precision at national level. A power allocation with power values between 0 and 1 is a tradeoff between the national level precision and the domain level precision. Since the sample size is usually large at the national level, the national level precision is not a concern.

In Table 1.3 below, we give an example of a proportional sample allocation of 15,000 individuals to 11 domains and to their urban-rural areas. The minimum domain sample size is 384 for domain 2, which is too small for estimating the total fertility rate (TFR) and childhood mortality rates. The largest sample size is for domain 11 which may be unnecessarily large. The actual total sample size given in the total row may be slightly different from the desired sample size because of rounding.

Table 1.3 Sample allocation: Proportional allocation

Serial Num	Total sample size =>	15000	Power value domain=>		Power value urban=>		
	Domain/Stratum Name/ID	Domain/ stratum size	Proportion urban	Sample Allocation			Specific Allocation	
				Urban	Rural	Domain	Urban	Rural
1	Domain 1	0.072	0.352	382	701	1083		
2	Domain 2	0.026	0.317	122	262	384		
3	Domain 3	0.070	0.568	597	454	1051		
4	Domain 4	0.142	0.275	586	1544	2130		
5	Domain 5	0.060	0.323	292	611	903		
6	Domain 6	0.046	0.135	92	593	685		
7	Domain 7	0.048	0.194	141	586	727		
8	Domain 8	0.094	0.251	354	1055	1409		
9	Domain 9	0.164	0.288	709	1749	2458		
10	Domain 10	0.091	0.191	262	1104	1366		
11	Domain 11	0.187	1.000	2803	0	2803		
Total		**1.000**	**0.423**	**6339**	**8660**	**14999**		

If we impose a condition such that the sample size should not be smaller than 1000 in each domain, after trying various power values, we find that a power value of 0.25 is appropriate, as shown in Table 1.4. In this case, we would have a minimum sample size of 1,022 for domain 2. Since domain 11 has only urban areas, the power allocation among the domains brought down the urban percentage in the sample. In order for urban areas to be properly represented, over sampling is applied in the urban areas of the other domains. With a power value of 0.65, the urban proportion in the sample is close to the proportion of the target population.

Table 1.4 Sample allocation: Power allocation

Serial Num	Total sample size =>	15000	Power value domain=>	0.25	Power value urban=>		0.65
	Domain/Stratum Name/ID	Domain/ stratum size	Proportion urban	Sample Allocation			Specific Allocation	
				Urban	Rural	Domain	Urban	Rural
1	Domain 1	0.072	0.352	533	791	1324		
2	Domain 2	0.026	0.317	386	636	1022		
3	Domain 3	0.070	0.568	716	599	1315		
4	Domain 4	0.142	0.275	546	1023	1569		
5	Domain 5	0.060	0.323	484	782	1266		
6	Domain 6	0.046	0.135	271	910	1181		
7	Domain 7	0.048	0.194	341	858	1199		
8	Domain 8	0.094	0.251	466	949	1415		
9	Domain 9	0.164	0.288	581	1045	1626		
10	Domain 10	0.091	0.191	395	1009	1404		
11	Domain 11	0.187	1.000	1680	0	1680		
Total		**1.000**	**0.423**	**6399**	**8602**	**15001**		

In Table 1.4, the small domains are oversampled compared with a proportional allocation. Oversampling some small domains is frequently practiced if domain level precision is required.

However, oversampling a small domain too much will harm the precision at national level. To prevent this, it is recommended to regroup the small domains to form domains of moderate size, especially when there is a very unequal population distribution among geographic domains, however, this is sometimes not possible due to political considerations.

The above discussion also applies to sample size allocation to strata within a domain where the domain sample size is fixed. A proportional allocation with sample size proportional to stratum size is good for all indicators and provides the best precision for the domain as a whole.

1.8 Two-stage cluster sampling procedure

The MEASURE DHS program utilizes a convenient and practical sample selection procedure for household based surveys developed on the basis of experience from past surveys—a *two-stage cluster sampling* procedure. A *cluster* is a group of adjacent households which serves as the PSU for field work efficiency. Interviewing a certain number of households in the same cluster can reduce greatly the amount of travel and time needed during data collection. In most cases, a cluster is an EA with a measure of size equal to the number of households or the population in the EA, provided by the population census.

At the first stage, a stratified sample of EAs is selected with *probability proportional to size* (PPS): in each stratum, a sample of a predetermined number of EAs is selected independently with probability proportional to the EA's measure of size. In the selected EAs, a listing procedure is performed such that all dwellings/households are listed. This procedure is important for correcting errors existing in the sampling frame, and it provides a sampling frame for household selection.

At the second stage, after a complete household listing is conducted in each of the selected EAs, a fixed (or variable) number of households is selected by equal probability *systematic sampling* in the selected EAs. In each selected household, a household questionnaire is completed to identify women age 15-49, men age 15-59 (15-54 or 15-49 in some surveys) and children under age five. Every eligible woman will be interviewed with an individual questionnaire, and every eligible man will be interviewed with an individual men's questionnaire in those households selected for the men's interview.

The advantages of this two-stage cluster sampling procedure can be summarized as follows:

1) It guarantees a representative sample of the target population when a list of all target individuals is not available which prohibits a direct sampling of target individuals;

2) A household listing procedure after the selection of the first stage and before the main survey provides a sampling frame for household selection in the central office;

3) The use of residential households as the second-stage sampling unit guarantees the best coverage of the target population; and

4) It reduces unnecessary sampling errors by avoiding more than two stages of selection (which usually uses a large PSU in the first stage of selection).

See more details in Sections 1.10 and 1.11 on household listing and selection, Chapter 2 on household listing, and Sections 3.2 and 3.3 of Chapter 3 on systematic sampling and sampling with probability proportional to size (PPS).

1.9 Sample “take” per cluster

Once the total sample size is determined and allocated to different survey domains/strata, it should be decided how many individuals (sample take) should be interviewed per sample cluster and then convert the domain/stratum sample size to number of clusters. Since the survey cost can be very different across the survey domains/strata, the sample take can have a big influence on the total survey budget. With a fixed sample size, a small sample take is good for survey precision because of the reduction of the design effect, but is expensive because more clusters are needed. The number of clusters affects the survey budget more than the overall sample size due to the travel between clusters during data collection, which represents an important part of field costs in rural areas. The MEASURE DHS program proposes a sample “take” of about 25-30 women per rural cluster. In urban areas, the cost advantage of a large “take” is generally smaller, and MEASURE DHS recommends a “take” of about 20-25 women per urban cluster. Since in most DHS surveys, the number of eligible women age 15-49 is very close to one per household, the sample take of individuals is equivalent to the sample take of households; therefore, in the following sections we refer to the sample take (or cluster take) as the number of sample households per cluster.

1.9.1 Optimum sample take

The optimum number of households to be selected per cluster depends on the variable under consideration, the intracluster correlation ρ, and the survey cost ratio c_1/c_2, where c_1 represents the cost per cluster including mainly the cost associated with travelling between the clusters for survey implementation (household listing and interview); while c_2 represents the cost per individual interview (the interviewing cost) and other costs of doing fieldwork within a cluster. A larger sample take per cluster and fewer clusters reduces survey field costs if the cost ratio is high, but it could also reduce the survey precision if the intracluster correlation is strong.

The MEASURE DHS Program has accumulated information on sampling errors for selected variables for many surveys throughout the world. Using this information, Aliaga and Ren (2006) conducted a research study to determine the optimum sample take per cluster. The results of the study have informed current practice in DHS surveys. If the average cluster size is around 250 households, a sample take of 20-30 households per cluster is within the acceptable range in most surveys. The research also supports the practice of setting a larger sample take in rural clusters than in urban clusters. Usually, the cost ratio in urban areas is smaller than that in rural areas. This would lead to a smaller sample take in an urban cluster than in a rural cluster. In sum, this research indicates that for the most important survey indicators, a sample take between 20 to 25 households is appropriate in urban clusters and a sample take between 25 to 30 households is appropriate in rural clusters.

Based on values of c_1/c_2 and ρ obtained from eight surveys, Table 1.5 below shows optimal sample takes for the indicator “proportion of currently married women 15-49 currently using any contraceptive method.” This indicator has a moderate intracluster correlation relative to other important survey indicators.

Table 1.5 Optimal sample take for currently married women 15-49 currently using any contraceptive method based on intracluster correlation ρ and survey cost ratio c_1/c_2 from past surveys

Country	Survey cost ratio c_1/c_2	Intracluster correlation ρ	Optimal sample take
Country 1	10	0.025	20
Country 2	10	0.037	16
Country 3	12	0.067	13
Country 4	12	0.052	15
Country 5	15	0.084	13
Country 6	27	0.031	29
Country 7	48	0.058	28
Country 8	52	0.023	47
Average	**23**	**0.047**	**23**

1.9.2 Variable sample take for self-weighting

A fixed sample take per cluster is easy for survey management and implementation, but it requires sampling weights that vary within a stratum. Different sampling weights result in larger sampling errors compared with a similar sample of constant weight within a sampling stratum, i.e., a *self-weighting sample*. A self-weighting sample consists of a sample of individuals in which each individual has the same probability of being selected, and therefore a constant sampling weight is used. In some cases a self-weighting sample is preferred for various reasons:

- it is equally representative for every individual of the target population;
- it reduces sampling errors.

Since the sample for DHS surveys is usually the result of a two-stage cluster sampling design, it is necessary to coordinate the sample take for each of the selected clusters. In an overall self-weighting sample, every individual in the target population has an equal probability of selection, which results in a proportional allocation. However, proportional allocation is not feasible when sampling domains are very different in size. Self-weighting at domain/stratum level, by contrast, is easy to achieve.

Let *n* be the total number of clusters selected for a DHS survey, let n_h be the number of clusters allocated to the h^{th} stratum; let X_h be the total number of households in the stratum *h*, let x_{hk} be the number of households in cluster *k* of stratum *h*, given by the sampling frame; then the selection probability of cluster *k* in stratum *h* is given by:

$$\pi_{hk} = \frac{n_h x_{hk}}{X_h}$$

Let x'_{hk} be the number of households listed in the cluster in the household listing operation, let m_h be the number of households to be selected from the cluster for a fixed sample take, then the overall selection probability of a household in the cluster is given by:

$$f_{hk}=\pi_{hk}\times\frac{m_h}{x'_{hk}}=\frac{n_h x_{hk}}{X_h}\times\frac{m_h}{x'_{hk}}$$

If $x'_{hk}=x_{hk}$ exactly for all *k* in stratum *h*, then it is easy to see that self-weighting is achieved in stratum *h* by a constant sample take m_h in all clusters since $f_h=\frac{n_h m_h}{X_h}$ is a constant in stratum *h*.

In practice, it is not possible that $x'_{hk}=x_{hk}$ for all *h* and *k*, especially when the last population census is no longer new. Therefore there is a need for sample coordination in order to achieve self-weighting. Let f_h and m_h be the calculated sampling fraction and average sample take in stratum *h* according to the sample allocation with $m_h=\frac{f_h X_h}{n_h}$; the number of households needed to achieve self-weighting in cluster *k* of stratum *h* is given by

$$m_{hk}=\frac{f_h X_h}{n_h}\times\frac{x'_{hk}}{x_{hk}}=m_h\times\frac{x'_{hk}}{x_{hk}}$$

which is a function of the ratio of the number of households listed over the number of households given in the sampling frame for every cluster: take more if more are listed or take fewer if fewer are listed. The above formula also shows that the sampling fraction is not a necessary parameter for sample take calculation. Using the designed average sample take is a more direct method because the sampling fraction is an abstract number. This formula is used in the self-weighting household selection templates presented in Chapter 3, Section 3.2. The relationship between the sample take and the cluster selection probability is given by

$$m_{hk}=\frac{f_h x'_{hk}}{\pi_{hk}}$$

For practical considerations, the sample take calculated above needs to be adjusted if is it too small or too large. Usually, we apply a cut-off to control the sample take within the range of a minimum of 10 households and a maximum of 50 households per cluster. For the clusters where the cut-off is applied, the sample is no longer self-weighting.

The advantages and disadvantages of a self-weighting sample can be summarized as:

Advantages:

1) Equally representative for every individual within a sampling stratum.

2) Reduced sampling errors.

Disadvantages:

1) Difficult for survey management (for example, to distribute the work-load) because of the variant sample take by cluster.

2) Difficult to control the expected sample size because of possible cut-offs, especially when the upper limit cut-offs are employed.

3) The self-weighting is not exact because of the rounding of the sample takes and this will bring bias in the survey estimation.

4) Self-weighting at the national level will break down the specific sample allocation at the domain/stratum level and bring the sample allocation back to a proportional allocation.

It is possible to overcome the second and the third disadvantages through a recursive calculation of sample take by re-distributing the cut-offs to the rest of the clusters in the stratum or control area, and by using a randomized sample take which allows non-integer numbers as sample size. Excel templates for both the traditional procedure and revised procedure are available.

1.10 Household listing

The household listing operation is a fundamental operation in DHS surveys. After the EAs are selected for the survey, a complete listing of dwelling units/households in the selected EAs is conducted prior to the selection of households. The listing operation consists of visiting each of the selected clusters, collecting geographic coordinates of the cluster, drawing a location map of the cluster as well as a sketch map of the structures in the cluster, recording on listing forms a description of every structure together with the names of the heads of the households in the structures and other characteristics. Mapping and listing of households represents a significant field cost, but it is essential to guarantee the exactness of sample implementation.

The listing operation is an important procedure for reducing non-sampling errors in the survey, especially when the sampling frame is outdated. The listing operation provides a complete list of occupied residential households in the EA. This information is necessary for an equal probability random selection of households in the second stage. With the household listing prior to the main survey, it is possible to pre-select the sample households in advance and the interviewers are asked to interview only the pre-selected households without replacement of non-responding households. With the sketch map and the household listing of the cluster produced in the household listing operation, the sampled households can be easily relocated by interviewers later. The fieldwork procedure for DHS surveys is designed to be replicable and therefore allows easy supervision; all these elements are designed to prevent serious bias during data collection.

It is sometimes suggested that listing could be avoided by making segments so small that they are equal to the required sample “take” per cluster. One could then use a “take-all” rule at the last stage of sampling. Such small segments, however, will generally be difficult to delineate. In planned urban areas, this difficulty may be reduced—one could adopt blocks, or even single buildings, as segments—but urban units of this kind are likely to be homogeneous, containing similar households, and therefore less than ideal as sampling clusters.

It is also not acceptable to attempt to avoid listing altogether by having interviewers create clusters as they go along, or by selecting the sample households at fixed intervals during a random walk up to a predetermined quota. Such methods are not acceptable because first, they do not guarantee a nonzero probability to every potential respondent; second, the procedure is not replicable, which complicates the field work supervision; and third, it can end up with a sample of *easy units* because of the lack of effort to make call backs to households or individuals who were not available at the first attempt to interview.

Listing costs can be reduced by using segmentation to decrease the size of the area which has to be listed; however, segmentation generates its own costs, and skill in map making and map

interpretation is required. Segmentation becomes progressively more difficult as segments become smaller because there are not enough natural boundaries to delineate very small segments. Moreover, concentration of the sample into smaller segments increases the sampling error. Since neighbors' characteristics are correlated, a smaller segment captures less of the variety existing in the population; this leads to less efficient sampling. There is a point beyond which it is not useful to attempt further segmentation. As a general rule the average segment size should not be less than 500 in population (approximately 100 households) in both urban and rural areas. However, segmentation has less economical effect in urban areas because the urban EAs are in general small geographic areas.

It is quite probable that some traditional tools in the household listing process will be modified in the future by using more sophisticated technology such as the *geographic positioning systems* (GPS) in order to collect more precise location information for the selected EAs. With this new tool we can produce more precise distribution maps of the structures with less supervision than in the traditional approach. The main feature is that every selected EA and every selected structure/dwelling can be located with high precision and thus relocated later, if desirable. In addition, GPS information is used more and more in DHS data analysis and presentation. At present, though, the recommended protocol for collecting GIS information in DHS surveys is to collect one coordinate for every selected cluster. See Chapter 2 for more details of the household listing operation.

1.11 Household selection in the central office

After the household listing operation, once the central office receives the completed listing materials for a cluster, they must first create a serial number for each of the occupied residential households, beginning with 1 and continuing to the total number of occupied residential households listed in the cluster. An occupied residential household designates those households occupied at the time of the listing, even if the occupant refused to cooperate at the time of listing, and those households where the occupants were absent at the time of listing but neighbors confirmed that they would not be absent for a long period and would be at home during the period of the main survey. Only occupied residential households should be numbered. This serial number is an ID number for the households. The household selection procedure will be performed based on this serial number. Whether or not a household is considered occupied at the time of the listing is very important because this fact will be related to the proportion of vacant households in the main survey.

The MEASURE DHS program has used several methods[3] for selecting households within clusters including:

1) Systematic selection: From a random starting point select every nth household (see Chapter 3 Section 3.2 for more details).

2) Systematic selection with runs: From a random starting point, select a group of sequential households called a "run". Several runs may be used within a cluster. Runs are selected with systematic selection. Selecting households in runs can greatly reduce the amount of travel within cluster during data collection, especially in rural clusters where households can be far apart.

The advantages of household selection in the central office can be summarized as:

[3] The MEASURE DHS program has developed various Excel templates for household selection in the central office: systematic selection, systematic selection with runs, self-weighting selection with and without control of sample size and with or without runs. Once the household listing is completed, it is possible to just copy the number of households listed in a cluster into the spreadsheet and the spreadsheet will show the selected household numbers automatically. See Chapter 3 Section 3.2.2 for details.

1) It allows for a check of coverage of the household listing results before the main survey and for the review and possible relisting of problematic clusters in advance.

2) Sampled households are pre-determined which prevents potential bias introduced by allowing the interviewers to select in the field which households are to be interviewed.

3) The field work procedure is exactly replicable which provides the possibility of easy and close supervision of the field work.

4) It is easier to control the work load for each interviewing team.

However, in cases when travelling between clusters represents a substantial cost, it is possible to forego the step of selecting households in the central office. In such cases, the household listing operation and the main survey can be combined into a single field operation. No essential changes are needed in the household listing procedure or household numbering, but making a detailed sketch map for the cluster may not be necessary because the listing team and the interviewing team are the same, and the household interview will begin immediately after the listing, so identifying the exact selected households during a separate visit is no longer a problem. The household selection must be done in the field manually if portable computers are not available. Some manual selection procedures have been developed for this purpose. Household listing and interviewing are two very different jobs, so in surveys where listing, selection and interviewing takes place in the same visit by the same staff, it may be necessary to conduct more extensive training of field teams before the field work begins and to supervise the teams more closely during the fieldwork. See Chapter 3 Section 3.2.2 for more details for manual household selection.

1.12 Household interviews

The household interview procedure is out of the scope of this manual since it is explained in detail in the interviewer's manual. This section will briefly discuss the main statistical points of the household interview. After the household selection, interviewers will be recruited and trained for the household and individual interviews. The training of the interviewer is an intensive training lasting at least four weeks for a standard DHS survey, and longer if the survey includes many biomarkers. Prior to the training, a pretest of the questionnaire will be conducted in a small number of clusters not selected for the main survey to assess the quality of the questionnaires and the understanding of the translations by interviewers and respondents. Problems and potential errors observed in the pretest will be addressed and resolved prior to fieldwork training. Finally, the interviewing team will be sent to selected clusters with a certain work load per team.

Once training is complete, teams of interviewers will be assigned a list of clusters and deployed to the field. Upon arrival in a new area, the interviewer team must first contact the local authorities for help to identify the correct cluster and to solicit cooperation during the field work. A team leader or supervisor is assigned for each interviewing team. The supervisor is responsible for cluster identification and should guarantee that the correct cluster will be interviewed. After checking the listing materials and verifying with the local authorities, the supervisor will distribute the sampled households among the interviewers. After locating a selected household, the interviewer will begin with a brief household interview, listing household members and visitors, and identifying among them all eligible women and men for the individual interview. Eligible individuals are defined as those who are in the specified age group (15-49), and are either usual members of the selected household or who slept in the household the night before the interviewer's visit.

Conscious omission of eligible individuals on the part of an interviewer by mis-reporting their age outside of the eligible age group is a real concern. Measures to eliminate this problem should be

undertaken. For example, the field editor should check the consistency of each completed questionnaire and, if suspicious things are identified, should return to the household for further verification of key items such as the number of household members, number of eligible individuals and number of children under age five.

In the event of failure to contact a household or an eligible person in the first visit, the interviewer is required to make at least two repeat visits, or call backs, on different days and at different times of the day before the interview is abandoned. The process of making call backs requires the teams to stay in a cluster for at least two to three days. Some countries propose large interviewing teams in order to try to cover an entire cluster in one day. This process is not acceptable for a DHS survey, even when the designed sample size can bear a large non-response rate, because non-response biases the survey results. A quick survey usually ends up with poor data quality. Both theory and practice prove that call backs and efforts to get difficult units to respond to the survey are the best way to remove bias and reduce the non-sampling errors to a minimum. For more details, refer to the DHS Survey Organization Manual and the Interviewer's Manual.

1.13 Sampling weight calculation

1.13.1 Why we need to weight the survey data

A DHS sample is a representative sample randomly selected from the target population. Each interviewed unit (household and individual) represents a certain number of similar units in the target population. In order for any statistical inferences drawn from the survey data to be valid, this representativeness of the sample must be taken into account. In general terms, sampling weights are used to make the sample more like the target population. All analyses should use the sampling weights calculated for each interviewed household and for each interviewed individual.

A sampling weight is an inflation factor which extrapolates the sample to the target population. For example, if equal probability sampling (or a self-weighting sample) is applied in a domain with a sampling fraction 1/500, this means that each sampled individual represents 500 similar individuals in the target population. Therefore, if we observed one particular individual having secondary education, we would conclude that there are 500 individuals in the target population having secondary education, corresponding to this particular individual. The total number of individuals with secondary education in the target population would be 500 times the total number of interviewed individuals having secondary education observed in the sample. This explanation also applies to unequal probability sampling. It is very important that sampling weights are properly calculated and applied in data analysis; otherwise, serious bias may be introduced, leading to incorrect conclusions.

Although all of the DHS indicators are means, proportions, rates or ratios, since a nationwide self-weighting sample is not usually feasible due to study domains as explained in Section 1.9, sampling weights are always necessary. Even when a survey is designed to be nationally self-weighting, it is necessary to correct for the different response patterns across domains/strata (see Section 1.13.4 for more details). Therefore, even surveys with self-weighting sample designs require the use of sampling weights.

Though the effect of sampling weights on survey indicators may be small, it is necessary to use sampling weights for the following reasons:

1) For valid statistical inference.

2) For correcting or reducing bias; weighting can reduce bias introduced by non-response or other non-sampling errors.

3) For keeping the weighted sample distribution close to the target population distribution, especially when oversampling is applied in certain domains/strata.

1.13.2 Design weights and sampling weights

The MEASURE DHS program calculates both *design weights* and *sampling weights* (or *survey weights*) for both households and individuals. The *design weight* of a sampling unit (household or individual) is the inverse of the overall probability with which the unit was selected in the sample. The *sampling weight* of a sampling unit is the design weight corrected for non-response or other calibrations.

Since is the DHS protocol involves no selection of eligible individuals within a sampled household (except for the domestic violence module, in which one eligible woman is selected from a sampled household), all eligible individuals from the same household share the same design weight, which is the same as the household's design weight. Therefore, the design weight is the basic weight for DHS surveys. All other weights are calculated based on the design weight. In calculating the sampling weight, it is possible to correct for both *unit non-response* (a sampling unit is not interviewed at all) and *item non-response* (the sampling unit does not provide answer for a specific question). The policy of the MEASURE DHS program is to correct for unit non-response at the stratum level (see Section 1.13.4) and leave the correction of item non-response to data users because it is variable specific. Correction of unit non-response at cluster level will increase the variability of sampling weights and therefore increase sampling errors. Because the correction for unit non-response is the same for an entire cluster and because household selection within a cluster is an equal probability selection, all the households in the same cluster share the same design weight and sampling weight, and the same is true for all individuals in the same cluster. This means that the DHS weights (both design weights and sampling weights) are cluster weights.

1.13.3 How to calculate the design weights

Assuming that a DHS survey sample is drawn with two-stage, stratified cluster sampling, design weights will be calculated based on the separate sampling probabilities for each sampling stage and for each cluster. We use the following notations:

P_{1hi}: first-stage sampling probability of the i^{th} cluster in stratum h
P_{2hi}: second-stage sampling probability within the i^{th} cluster (household selection)

Let n_h be the number of clusters selected in stratum h; let M_{hi} be the measure of size of the cluster used in the first stage's selection, usually the measure of size is the number of households residing in the cluster according to the sampling frame; let $\sum M_{hi}$ be the total measure of size in the stratum h. The probability of selecting the i^{th} cluster in the sample is calculated as follows:

$$P_{1hi}=\frac{n_h M_{hi}}{\sum M_{hi}}$$

Let b_{hi} be the proportion of households in the selected cluster compared to the total number of households in EA i in stratum h if the EA is segmented, otherwise $b_{hi}=1$. Then the probability of selecting cluster i in the sample is:

$$P_{1hi} = \frac{n_h M_{hi}}{\sum M_{hi}} \times b_{hi}$$

Let L_{hi} be the number of households listed in the household listing operation in cluster *i* in stratum *h*; let t_{hi} be the number of households selected in the cluster. The second stage selection probability for each household in the cluster is calculated as follows:

$$P_{2hi}=\frac{t_{hi}}{L_{hi}}$$

The overall selection probability of each household in cluster *i* of stratum *h* is therefore the product of the selection probabilities of the two stages:

$$P_{hi}=P_{1hi}\times P_{2hi}$$

The design weight for each household in cluster *i* of stratum *h* is the inverse of its overall selection probability:

$$d_{hi}=1/P_{hi}$$

The calculation of the design weight is not complicated; however, difficulties often result from not having of all the design parameters involved in the above calculation because they are not well documented, especially when the sampling frame is a master sample. See Chapter 5 for more details on sample documentation.

1.13.4 Correction of unit non-response and calculation of sampling weights

The design weight calculated above is based on sample design parameters. If there is no non-response at the cluster level, at the household level, or at the individual level, the design weight is enough for all analyses, for both household indicators and individual indicators. However, non-response is inevitable in all surveys, and different units have different response behaviors. The experience of the MEASURE DHS program shows that urban households are less likely to respond to the survey than their counterparts in rural areas, households in developed regions are less likely to respond to the survey than their counterparts in less-developed regions, rich households are less likely to respond to the survey than poor households, individuals with higher levels of education are less likely to respond to the survey than those with lower levels of education, men are less likely to respond to the survey than women, and so forth.

The idea of correcting for unit non-response is to calculate a response rate for each homogeneous response group, then inflate the design weight by dividing it by the response rate for each response group. The construction of homogeneous response groups depends on the knowledge of the response behavior of the sampling units. DHS surveys always use the sampling stratum as the response group because the stratification is usually achieved by regrouping homogeneous sampling units in a single stratum. It is possible to use a cluster as a response group, but the disadvantage is that the response rates may vary too much at the cluster level, which will increase the variability of the sampling weight; which in turn increases the sampling variance. Furthermore, correction of non-response at the cluster level will interfere with self-weighting if a self-weighting sample has been designed.

By assuming that the response groups coincide with the sampling strata, the following steps explain how to calculate the sampling weight by first calculating the various response rates for unit non-response. Please note that the response rates calculated here are different from the response rates calculated in Appendix A of DHS survey final reports. In Appendix A, household and individual

response rates are calculated as ratios of the number of interviewed units over the number of eligible units because the aim is just to show the results of survey implementation. Here we use weighted ratios because the aim is to correct the design weight to compensate for non-response, therefore the design weight should be involved. Because a non-responding unit with a large sampling weight will have a larger impact on survey estimates than a non-responding unit with a small design weight, a weighted response rate for correction of non-response is better than an un-weighted response rate.

1. Cluster level response rate

Let n_h be the number of clusters selected in stratum *h*; let $n_h^{'}$ be the number of clusters interviewed. The cluster level response rate in stratum *h* is therefore

$$R_{ch}=n_h^{'}/n_h$$

2. Household level response rate

Let m_{hi} be the number of households found (see Chapter 2, Section 2.10 for definition) in cluster *i* of stratum *h*; let $m_{hi}^{'}$ be the number of households interviewed in the cluster. The household response rate in stratum *h* is calculated by

$$R_{hh}=\sum d_{hi}m_{hi}^{'}/\sum d_{hi}m_{hi}$$

where d_{hi} is the design weight of cluster *i* in stratum *h*; the summation is over all clusters in the stratum *h*.

3. Individual response rate

Let k_{hi} be the number of eligible individuals found in cluster *i* of stratum *h*; let $k_{hi}^{'}$ be the number of individuals interviewed. The individual response rate in stratum *h* is calculated as

$$R_{ph}=\sum d_{hi}k_{hi}^{'}/\sum d_{hi}k_{hi}$$

where d_{hi} is the design weight of cluster *i* in stratum *h*; the summation is over all clusters in the stratum *h*.

The household sampling weight of cluster *i* in stratum *h* is calculated by dividing the household design weight by the product of the cluster response rate and the household response rate, for each of the sampling stratum:

$$D_{hi}=d_{hi}/(R_{ch}\times R_{hh}),$$ for cluster *i* of stratum *h*.

The individual sampling weight of cluster *i* in stratum *h* is calculated by dividing the household sampling weight by the individual response rate, or equivalently, by dividing the household design weight by the product of the cluster response rate, the household response rate and the individual response rate, for each of the sampling strata:

$$W_{hi}=D_{hi}/R_{ph}=d_{hi}/(R_{ch}\times R_{hh}\times R_{ph}),$$ for cluster *i* of stratum *h*.

It is easy to see that the difference between the household sampling weights and the individual sampling weights is introduced by individual non-response.

The sampling weights for households selected for the men's survey and for men can be calculated similarly. We need a separate household sampling weight for the men's survey in cases where the men's survey is conducted in a sub-sample of households selected for the women's survey, and we suppose that the response behavior of households in the men's survey sub-sample may be different from the overall household response rate.

If no normalization is requested, we can stop here. The above calculated household sampling weight and individual sampling weight can be used to produce any indicators at the household level

and the individual level, respectively. As we mentioned earlier in Section 1.13.1, a sampling weight is an inflation or extrapolation factor. The weighted sum of households interviewed

$$T=\sum\sum D_{hi}m_{hi}^{i}$$

is an unbiased estimate of the total number of ordinary residential households of the country; where m_{hi}^{i} is the number of households interviewed in the i^{th} cluster of stratum h, and the summation is over all clusters and strata in the total sample. Similarly, the weighted sum of all interviewed women

$$W=\sum\sum W_{hi}k_{hi}^{i}$$

is an unbiased estimate of the total women in the target population (women age 15-49) of the country; where k_{hi}^{i} is the number of women interviewed in the i^{th} cluster of stratum h, and the summation is over all clusters and strata in the total sample.

1.13.5 Normalization of sampling weights

Normalization of sampling weights is not necessary for survey data analysis. In order to prevent large numbers for the number of weighted cases in the tables in DHS survey final reports, it is the MEASURE DHS tradition to calculate *normalized standard weights* for both households and individuals. With the normalized standard weight, the number of unweighted cases coincides with the number of weighted cases at the national level for both total households and total individuals. The normalized standard weight of a sampling unit is calculated based on its sampling weight, by multiplying the sampling weight with a unique constant at the national level. The constant or the *normalization factor* is the total number of completed cases divided by the total number of weighted cases (based on the sampling weight). This number is equal to the estimated total sampling fraction because the total number of weighted cases with the sampling weight is an estimation of the total target population. Therefore the standard weights in the DHS data files are relative weights. Relative weights can be used to estimate means, proportions, rates and ratios because the normalization factor is cancelled out when used in both numerator and denominator, so it has no effect on the calculated indicator values. This point also explains why the normalization must be done at the national level and not the regional level: at the regional level, the normalization factor cannot be cancelled out, and bias will be introduced in the calculated indicator values. Because the normalized standard weights have no scale, they are not valid for estimating totals. Also the normalized weight is not valid for pooled data, even for data pooled for women and men in the same survey, because the normalization factor is country and sex specific.

1. Normalized household standard weight[4]

The normalization factor for calculating household standard weight is calculated as

$$FH=\sum\sum m_{hi}^{i}/\sum\sum D_{hi}m_{hi}^{i}$$

The household standard weight for cluster i in stratum h is calculated by

$$HV005_{hi}=D_{hi}\times FH=D_{hi}\times\sum\sum m_{hi}^{i}/\sum\sum D_{hi}m_{hi}^{i}$$

where HV005 is the household standard weight variable in the DHS Recode data files.

[4] The MEASURE DHS program has developed Excel templates for facilitating standard weight calculations. If all design parameters and the survey results (number of households found and interviewed, number of eligible women found and interviewed, number of eligible men found and interviewed, number of eligible women and men found and tested, by cluster) are provided in the input page, the standard weights will be calculated automatically in different pages.

It is easy to see that the weighted sum of households interviewed by using the standard weight equals the unweighted sum of households interviewed for the total sample. This condition will not be met at the domain level or for sub-populations. At the domain level, the weighted sum of households interviewed may be larger or smaller than the unweighted sum of households interviewed, depending on whether the domain is undersampled or oversampled.

2. Normalized women's standard weight

The normalization factor for calculating the women's standard weight is calculated as

$$FW = \sum\sum k_{hi}^{i} / \sum\sum W_{hi} k_{hi}^{i}$$

The women's standard weight for cluster i in stratum h is calculated by

$$V005_{hi} = W_{hi} \times FW = W_{hi} \times \sum\sum k_{hi}^{i} / \sum\sum W_{hi} k_{hi}^{i}$$

where V005 is the women's standard weight variable in the DHS Recode data files.

The standard weights for households selected for the men's survey and for men can be calculated in a similar way.

1.13.6 Standard weights for HIV testing

The sampling weights for HIV testing are calculated separately for women and men, but they are calculated using the same methodology. The only difference is in the calculation of the normalization factors, if a normalized weight is requested. In order to calculate the weighted HIV prevalence for women and men together using a normalized weight, the standard weight for HIV testing must be normalized for women and men together. In most DHS surveys, HIV testing is conducted in the same subsample of households selected for men's survey, and every woman or man in the household who is eligible for the individual interview is eligible for HIV testing. Once the household sampling weight for the men's survey is calculated using the procedures stated in Section 1.13.5, the sampling weights for HIV testing for women and men may be calculated separately by correcting the household sampling weight for the non-response rates of women and men for HIV testing, respectively. For simplicity, let MD_{hi} be the household sampling weight in cluster i of stratum h for the men's survey sub-sample, the response rates to HIV testing for women and men are calculated respectively by

$$WR_{hi} = \sum MD_{hi} WHIV_{hi}^{*} / \sum MD_{hi} WHIV_{hi}$$
$$MR_{hi} = \sum MD_{hi} MHIV_{hi}^{*} / \sum MD_{hi} MHIV_{hi}$$

where $WHIV_{hi}$ is the number of women eligible for HIV testing, and $WHIV_{hi}^{*}$ is the number of women tested with a valid test result, in cluster i of stratum h; $MHIV_{hi}$ and $MHIV_{hi}^{*}$ are the number of men eligible and the number of men tested with a valid test result, respectively, in cluster i of stratum h.

The sampling weights for HIV testing for women and men, respectively, are calculated by

$$HIV_{hi}^{W} = MD_{hi} / WR_{hi} \quad , \quad HIV_{hi}^{M} = MD_{hi} / MR_{hi}$$

In cluster *i* of stratum *h*, the normalized standard weights for HIV testing for women and men, respectively, are calculated by

$$HIV05_{hi}^{W} = HIV_{hi}^{W} \times \left(\sum\sum WHIV_{hi} + \sum\sum MHIV_{hi} \right) / \left(\sum\sum HIV_{hi}^{W} \times WHIV_{hi} + \sum\sum HIV_{hi}^{M} \times MHIV_{hi} \right)$$
$$HIV05_{hi}^{M} = HIV_{hi}^{M} \times \left(\sum\sum WHIV_{hi} + \sum\sum MHIV_{hi} \right) / \left(\sum\sum HIV_{hi}^{W} \times WHIV_{hi} + \sum\sum HIV_{hi}^{M} \times MHIV_{hi} \right)$$

where the double summations are over all clusters and strata in the total sample.

1.13.7 De-normalization of standard weights for pooled data

For all of the DHS data, the weight variables HV005 (household standard weight), V005 (women's standard weight) and MV005 (men's standard weight) are relative weights which are normalized so that the total number of weighted cases is equal to the total number of unweighted cases, for the three kinds of units. In some situations, such as analyses involving data from more than one survey, data users may need the un-normalized sampling weight for analyzing pooled data. As mentioned in Section 1.13.5, since normalization is country specific and sex specific, it is necessary to de-normalize the standard weights provided in the DHS Recode data files for analyzing pooled data.

The normalization procedure consists of multiplying the sampling weight by a normalization factor for the total sample. The normalization factor is the estimated total sampling fraction: the number of completed cases divided by the number of weighted cases by using the sampling weight, for each kind of sampling unit. The weighted number of cases with sampling weight is an estimation of the total target population. Therefore, in order to de-normalize a normalized weight, simply divide the normalized weight by the total sampling fraction. The estimated total sampling fraction is usually not provided in the DHS data file or in the final report. In order to calculate the total sampling fraction, it is necessary to know the total target population at the time of the survey. The total target population at the time of the survey is easy to get from various sources. The country's statistical office, the United Nations Population Division's (UNPD) World Population Prospects[5], and the United Nations Population Fund (UNFPA) are three sources that may be easy to access.

As mentioned above, if pooled data analysis is required, the standard weight variables HV005, V005 and MV005 must be rescaled or de-normalized. The de-normalization procedure is the inverse of the normalization procedure: that is, multiply the standard weight by the target population and divide by the number of completed cases, for each survey. The de-normalized weights for households, women and men (HV005*, V005*, and MV005*, respectively) can be calculated using the following formulas:

HV005* = HV005 × (total number of residential households in the country)/
(total number of households interviewed in the survey)

V005* = V005 × (total female population 15-49 in the country)/
(total number of women 15-49 interviewed in the survey)

MV005* = MV005 × (total male population 15-49 (15-59) in the country)/
(total number of men 15-49 (15-59) interviewed in the survey)

[5] http://esa.un.org/unpd/wpp/index.htm

If normalized weights are preferred, the above re-scaled weights can be re-normalized by multiplying by the total number of completed women's and men's interviews combined, dividing by the total number of weighted cases combined, and applying the above re-scaled weights to the pooled data.

Note that the normalization of sampling weights is done for the total sample for households, women and men separately. If the aim is to tabulate indicators for a certain sub-population from pooled data, for example, vaccination coverage for children 12-23 months, the de-normalization has nothing to do with the total population of children 12-23 months because there is no standard weight calculated for children 12-23 months in DHS surveys. If the indicator is tabulated at the household level using the household weight, the household standard weights must be de-normalized for all of the surveys included in the analysis as explained above; likewise, if the indicator is tabulated at the individual level using the women's (or child's mother's) weight, the women's standard weights must be de-normalized for each of the surveys.

1.14 Calibration of sampling weights in case of bias

Generalized calibration (Deville and Särndal, 1992; Deville et al, 1993) has now become a popular and powerful framework in survey data analysis for statistical offices in many countries. It allows for the utilization of different sources of auxiliary information to improve estimates from sample surveys. Calibration can reduce sampling errors, can correct bias caused by non-response and other non-sampling errors, and can reduce the influence of extreme values. Calibration is a "weight tuning" procedure such that the tuned sampling weight can produce estimates without error for known population characteristics. The precision of an estimator using a calibrated weight is equivalent to a regression estimator but is much easier to calculate with the help of calibration software such as CALMAR, a SAS Macro procedure developed by the French Institute of Statistics and Economic Studies (INSEE), and the SPSS procedure developed by Statistics Belgium. DHS surveys employ calibration of sampling weights only in cases where serious bias is observed in the collected data, and there is reliable auxiliary information available for the calibration.

Let X be a multivariate auxiliary variable with p components such that the population totals of each of the component variables are known beforehand from the recent population census, that is, $t_x = \sum_{i \in U} X_i = (t_{x_1}, t_{x_2}, \ldots, t_{x_p})^\tau$ is known. Let x_i be the observations of the auxiliary variables from the survey $x_i = (x_{1i}, x_{2i}, \ldots, x_{pi})^\tau$ for the respondent sampling unit i. Let D_i be the sampling weight for unit i. The calibration procedure consists of modifying the sampling weight slightly from D_i to W_i such that a given distance measure between the sampling weights D_i and the calibrated weights W_i

$$\sum_{i \in s} g(W_i, D_i)$$

is minimized under the constraints

$$\sum_{i \in s} w_i x_i = t_x$$

where g is a distance function which measures the distance between D_i and W_i. The constraints imposed are that the known auxiliary variable totals are estimated without error with the calibrated weights. If the variable of interest is well correlated with the auxiliary variables, then we expect that the precision can be greatly improved for estimating the variable of interest. The calibration theory states that the calibrated weights have the following formula

$$w_i = d_i F(q_i x_i' \lambda(s))$$

where $F(i)$ is called the calibration function which is the reciprocal of the derivative of the distance function g; q_i is a calibration weight which is usually set to 1 in the lack of prior knowledge; $\lambda(s)'$ is a constant depending on the particular sample s which is to be solved. When $F(x_i^T \lambda(s)) = (1 + q_i x_i^T \lambda(s)')$, which corresponds to one of the five proposed calibration functions in Deville et al, 1993, it is easy to solve, $\lambda(s)'$ is given by

$$\lambda(s) = T_s^{-1}(t_x - \hat{t}_{\pi x})$$

with

$$T_s = \sum_{i \in s} d_i q_i x_i x_i'$$

For a given variable of interest y, the calibrated estimator of the population total is equivalent to the generalized regression estimator

$$\hat{t}_y = \sum_{i \in s} w_i y_i = \hat{t}_{\pi y} + \hat{B}_s^T (t_x - \hat{t}_{\pi x})$$

where $\hat{B}_s = T_s^{-1} \sum_{i \in s} q_i D_i x_i y_i$ is the sample estimation of the regression coefficient; $\hat{t}_{\pi y}$ and $\hat{t}_{\pi x}$ are the simple estimators using the sampling weight

$$\hat{t}_{\pi y} = \sum_{i \in s} D_i y_i \quad , \quad \hat{t}_{\pi x} = \sum_{i \in s} D_i x_i$$

A mean estimation of the variable of interest y can be calculated by

$$\bar{\hat{Y}} = \frac{\sum_{i \in s} w_i y_i}{\sum_{i \in s} w_i}$$

The calibration estimator can be equivalently formulated with known proportions of one or more auxiliary variables. The calibration can be conducted at the individual level, which will result in an individual specific weight, or it can be conducted at the cluster level with aggregated data, which will result in a cluster weight. For more details see the related references given in the end of this document.

1.15 Data quality and sampling error reporting

Data quality is always a major concern for all MEASURE DHS projects. Though numerous efforts are made in implementing DHS surveys to maximize the quality of the data collected, non-sampling errors are always the main concerns for data quality. Data quality of a survey directly affects the reliability of the statistics produced. Many countries have laws that require reports of survey findings to include an evaluation of data quality and reliability. Data quality can be measured by total survey error including bias introduced by various sampling and non-sampling errors.

DHS survey final reports usually include tables in an appendix for data quality evaluation purposes, including: age distributions of household population by sex; age distributions of eligible and interviewed women and men; completeness of reporting on date of birth, age at death, age/date at first union, education and anthropometric measures, etc. The MEASURE DHS program also conducts some in-depth studies on data quality for specific topics, which are provided in published reports.

Apart from the data quality tables, DHS survey final reports provide sampling errors for selected indicators in Appendix B. Sampling errors are important reliability measures which tell the user the degree of error associated with a particular estimated indicator value, the number of cases

involved in the calculation of the indicator, the efficiency or clustering effects of the sample design compared to a simple random sampling and the range for the true value of an indicator at a certain confidence level. The reader is referred to Chapter 4, Section 4.2 for more details on sampling errors and their calculation.

DHS survey final reports also provide an appendix on the sample design of the survey. The sample design document reports the survey methodology used for the survey, including the aim of the survey, the target population, the sample size, the reporting domains, the stratification and sample allocation, sample selection procedure, sampling weight calculation, correction for non-response, calibration of sampling weights, and the results of survey implementation. See Chapter 5, Section 5.2 for more details on sample design.

1.16 Sample documentation

The task of a sampling statistician does not end with the selection of the sample. The preservation of sampling documentation is an essential requisite for sampling weight calculation, for sampling error computation, for data quality evaluation, for linkage with other data sources, and for various kinds of checks and supplementary studies. Special efforts are needed at the time of the sample design, at the end of the fieldwork, and at the completion of the data file if the task of sample documentation is to be carried out effectively. If preservation of documentation is delayed, considerable effort will be required to reconstitute the missing information when it is needed.

The sample documentation must comply with the survey confidentiality requirements. When HIV testing is conducted in a DHS or AIS (AIDS Indicator Survey), the confidentiality guidelines require the complete destruction of all intermediate documents which can potentially be used to identify any single household or individual who participated in the testing. This requirement reinforces the importance of timely sample documentation. See Chapter 5 for detailed requirements in sample documentation.

1.17 Confidentiality

The final data files for DHS surveys are made available to interested researchers. Therefore, the confidentiality of private information collected from individual respondents is a major concern, especially when sensitive information such as sexual activity and HIV status are collected. Protecting the confidentiality of the individual respondent is not only an ethical obligation, but it also promotes more accurate data because respondents are more likely to provide truthful responses if they feel confident their information will be kept private.

DHS surveys follow strict rules imposed at various steps during the survey implementation to prevent the direct or indirect disclosure of the identity of individual respondents. The principal pieces of information that can indirectly identify an individual respondent are cluster number, household number, the cluster selection probability and the sampling weights. The cluster number is an important identifier for sampling error calculations; the household number is important for household level and individual level data management and tabulation; the cluster selection probability is useful for cluster level modeling; and sampling weights are necessary for all analysis. So these variables must be present in the final data file. The household number in the final DHS data file is not informative, and sampling weights are not informative after correction of non-response and normalization. The cluster selection probability is potentially informative only if lower level identification information such as district and locality are present, and DHS survey final data files do not provide geographic information below the level of region or survey domain, especially when HIV testing is conducted. Thus the only concern is the disclosure of the cluster. For DHS or AIS surveys with HIV testing, the final data files provide scrambled cluster and household numbers for further insurance against disclosure.

2 HOUSEHOLD LISTING OPERATION

2.1 Introduction

DHS surveys are nationwide sample surveys designed to provide information on the levels of fertility, infant and child mortality, use of family planning, knowledge and attitudes toward HIV/AIDS and other sexually transmitted infections (STI), and on other family welfare and health indicators. The surveys generally interview women age 15-49 and men age 15-59 (15-49 or 15-54 in some surveys). The women and men to be interviewed live in ordinary residential households which are randomly selected from a set of sample points consisting of clusters of households. Prior to interviewing, all households located in the selected clusters will be listed. The listing of households for each cluster will be used in selecting the final sample of households to be included in the DHS survey.

The listing operation consists of visiting each cluster, recording on listing forms a description of every structure together with the names of the heads of the households found in the structure, and drawing a location map of the cluster as well as a detailed sketch map of all structures residing in the cluster. These materials will guide the interviewers to find the pre-selected households for interviewing and will allow field work supervisors to perform quality control during data collection.

The following sections present the general guidelines for conducting a household listing operation. Modifications may be needed to adapt to country specific situations.

2.2 Definition of terms

Following are brief definitions of the terms used in this document.

A census *Enumeration Area* (EA) is a geographical statistical unit created for a census and containing a certain number of households. An EA is usually a city block in urban areas and a village, a part of a village or a group of small villages in the rural areas with its location and boundaries well defined and recorded on census maps.

A *cluster* is the smallest geographical survey statistical unit for DHS surveys. It consists of a number of adjacent households in a geographical area. For DHS surveys, a cluster corresponds either to an EA or a segment of a large EA.

A *base map* is a reference map that describes the geographical location and boundaries of an EA.

A *structure* is a free-standing building or other construction that can have one or more dwelling units for residential or commercial use. Residential structures can have one or more dwelling units (for example: single house, apartment structure).

A *dwelling unit* is a room or a group of rooms normally intended as a residence for one household (for example: a single house, an apartment, a group of rooms in a house); a dwelling unit can also have more than one household.

A *household* consists of a person or a group of related or unrelated persons, who live together in the same dwelling unit, who acknowledge one adult male or female 15 years old or older as the head of the household, who share the same housekeeping arrangements, and are considered as one unit. In some cases one may find a group of people living together in the same house, but each person has separate eating arrangements; they should be counted as separate one-person households. Collective living arrangements such as army camps, boarding schools, or prisons will not be considered as households. Examples of households are:

- a man with his wife or his wives with or without children
- a man with his wife or his wives, his children and his parents
- a man with his wife or his wives, his married children living together for some social or economic reasons (the group recognize one person as household head)
- a widowed or divorced man or woman with or without children

The *head of household* is the person who is acknowledged as such by members of the household and who is usually responsible for the upkeep and maintenance of the household.

A *location map* is a map produced in the household listing operation which indicates the main access to a cluster, including main roads and main landmarks in the cluster. Sometimes it may be useful even to include some important landmarks in the neighboring cluster.

A *sketch map* is a map produced in household listing operation, with location or marks of all structures found in the listing operation which helps the interviewer to relocate the selected households. A sketch map also contains the cluster identification information, location information, access information, principal physical features and landmarks such as mountains, rivers, roads and electric poles.

2.3 Responsibilities of the listing staff

Persons recruited to participate in the household listing operation will work in teams consisting of two enumerators. A coordinator will monitor the entire operation.

The responsibilities of the coordinator are to:

1) obtain base maps for all the clusters included in the survey;
2) arrange for the reproduction of all listing materials (listing manuals, mapping and listing forms); the map information forms and the household listing forms must be prepared in sufficient numbers to cover all of the clusters to be visited.
3) assign teams to clusters;
4) monitor the reception of the completed listing forms at the central office; and
5) verify that the quality of work is acceptable.

If GPS coordinates are being collected during the listing operation, the coordinator must also:

6) obtain one GPS receiver per listing team, plus two backup receivers, and tag each GPS receiver with a number;
7) ensure that all GPS receivers have the correct settings (see Section 2.6 below) and distribute a receiver to each field team;
8) obtain and copy all GPS training materials for listing staff; and
9) train all listing staff to record GPS waypoints in the GPS units as well as on Form DHS/1.

The responsibilities of the enumerators are to:

1) identify the boundaries of the cluster;

2) draw a location map showing the location of the cluster;

3) draw a detailed sketch map of the cluster showing the locations of all structures residing in the cluster;

4) list all the households in the cluster in a systematic manner;

5) communicate to the coordinator problems encountered in the field and follow his instructions.

6) transfer the completed listing forms to the coordinator or to the central office;

If GPS coordinates are being collected during the listing operation, enumerators must also:

7) capture and record the GPS waypoint of the center of the cluster; and

8) complete the portion of form DHS/1 designated for GPS information for each cluster.

The two enumerators in each team should work together at the same time in the same area. They will first identify the cluster boundaries together. Then one enumerator prepares the location and the sketch map while the other does the household listing. The materials needed for the household listing operation are:

- Manual for Household Listing
- Base map of the area containing the cluster
- Map Information Form (Form DHS/1)
- Household Listing Form (Form DHS/2)
- Segmentation form (Form DHS/3)

If GPS coordinates are to be recorded during the listing operation, the following additional materials are needed:

- GPS receivers, batteries and cables
- GPS training manuals and handouts

2.4 Locating the cluster

The coordinator will provide the listing team with a base map containing the cluster assigned to the team. The listing team will typically make two tours of the cluster: the first to identify the cluster boundaries and to create the location map, and the second to create the listing and draw the sketch map. Upon arrival in a cluster, the team should first contact the local authorities for help in identifying the boundaries and get general information on the cluster, for example, the rough number of residential households in the cluster. In most cases, the cluster boundaries follow easily recognizable natural features such as streams or rivers, and construction features such as roads or railroads. In some cases, the boundaries may not be marked with visible features (especially in rural areas), attention should be paid to locate the cluster boundaries as precisely as possible according to the detailed description of the cluster and its base map.

Before doing the listing, the team should tour the cluster to determine an efficient route of travel for listing all of the structures. The cluster should be divided into parts if possible. A part can be a block of structures. The listing team will make a location map of the cluster indicating the boundaries of the parts, as well as the relative location of landmarks, public structures (e.g., schools, religious structures, public offices and markets) and main roads. This location map will serve as a guide for the interviewing team when they begin data collection.

2.5 Preparing location and sketch maps

The coordinator will designate one enumerator of the team as the *mapper*. The second enumerator will be the *lister*. Although the two have separate tasks to perform, they must move together and work in close cooperation; the mapper prepares the maps, and the lister collects information on the structures (and corresponding households) indicated on the sketch map.

The mapping of the cluster and the listing of the households should be done in a systematic manner so that there are no omissions or duplications. If the cluster consists of a number of blocks, then the team should finish each block before going to the next adjacent block. Within each block, start at one corner of the block and move clockwise around it. In rural areas where structures are frequently found in small groups, the team should work in one group of structures at a time and in each group they can start at the centre (choosing any landmark, such as a school, to be the centre) and move around it clockwise.

In the first tour of the cluster, the mapper will prepare a location map of the cluster on the Map Information Form (Form DHS/1). First, fill in the identification box for the cluster on the first page. All information needed for filling in the identification box is provided by the coordinator. In the space provided on the second page, draw a map showing the location of the cluster and include instructions on how to get to the cluster. Include all useful information to find the cluster and its boundaries directly on the map and in the space reserved for observations if necessary.

In the second tour of the cluster, using the third page of the Map Information Form, the mapper will draw a sketch map of all structures found in the cluster, including vacant structures and structures under construction. It is important that the mapper and lister work together and coordinate their activities, since the structure numbers that the mapper indicates on the sketch map must correspond to the serial numbers assigned by the lister on the listing form for the same structures.

On the sketch map, mark the starting point with a large X. Place a small square at the spot where each structure in the cluster is located. For any non-residential structure, identify its use (for example, a store or factory). Number all structures in sequential order beginning with "1". Whenever there is a break in the numbering of structures (for example, when moving from one block to another), use an arrow to indicate how the numbers proceed from one set of structures to another. Although it may be difficult to pinpoint the exact location of the structure on the map, even an approximate location is useful for finding the structure in the future. Add to the sketch map all landmarks (such as a park), public structures (such as a school or church), and streets or roads. Sometimes it is useful to add to the sketch map landmarks that are found outside the cluster boundaries, if they are helpful in identifying other structures inside the cluster.

Use the marker or chalk provided to write on the entrance to the structure the number that has been assigned to the structure. Remember that this is the serial number of the structure as assigned on the household listing form, which is the same as the number indicated on the sketch map. In order to distinguish the number from other numbers that may exist already on the door of the structure, write “DHS” in front of the number, for example, for the structure number 5, write “DHS/5,” similarly on the door of structure number 44 write “DHS/44.”

A structure is called a *multi-unit structure* if it contains more than one household in the structure. Otherwise it is called a single-unit structure. All households found in a structure or multi-unit structure must be numbered from 1 to m, within the structure[6]. The structure number plus the household number form a unique identification number for a household, and for all of the households in the cluster. For example, household number 3 in structure number 44 would be uniquely identified with ID number DHS/44-3. It is very useful to write the household ID number at the entrance of the household to later assist the interviewer to identify the household for interview.

2.6 Collecting a GPS waypoint for each cluster

A GPS waypoint is a latitude and longitude reading that represents a location. For some surveys, GPS data for EAs are available from the census. However, if the data are not available, or are of questionable quality, one GPS waypoint for each cluster should be recorded during the listing phase of the survey. These waypoints are recorded using a GPS unit (a Garmin ETREX unit is used in this guide) and data collection forms. If GPS units other than the Garmin ETREX are used, this guide will still be useful; however, some of the instructions may not apply due to differences in design and menus. The Garmin ETREX owner's manual may be useful to consult on the basics of the GPS unit.

Take one reading for each cluster. The GPS waypoints will be captured by the mapper while he is mapping the clusters. One GPS waypoint must be taken for each cluster, and in the case of large clusters which are being segmented, one point should be taken for each segment selected for listing. In DHS surveys, clusters are usually census EAs, sometimes villages in rural areas or city blocks in urban areas. Collecting only one waypoint for the cluster greatly reduces the chance of compromising confidentiality of the respondents and at the same time is sufficient to allow for the integration of multiple datasets for further analysis. The DHS cluster waypoint should always be taken at the geographic center of the cluster or segment. If the cluster is segmented, the point should be taken for the segment chosen by the Mapping and Listing Coordinator to be included in the survey.

***Save the waypoint and record the latitude, longitude, and altitude*.** The latitude, longitude, and altitude reading for a location are stored in two places: in the GPS unit's memory and on the DHS/1 paper form. GPS units can be broken or lost, and experience has shown that a hardcopy backup is essential. In addition, the paper form provides a backup should the data in the GPS unit be changed, deleted, or misidentified (i.e., the operator names the cluster incorrectly in the unit). Each position saved in the GPS unit is called a waypoint, and each waypoint has a unique name. If possible, the waypoint ID should be the same as the DHS cluster number. If it is not possible, the waypoint ID should be unique to the cluster and recorded on Form DHS/1 (do not record the same waypoint ID for two different clusters). When a waypoint is saved, the GPS unit assigns it a default name. The mapper must edit the default name and change it to the 6-digit DHS cluster ID number. For example, the waypoint for DHS cluster 101 would be named "000101". Cluster 1101 would be named "001101". After saving the waypoint, the mapper will use the identification box of the Map Information Form (Form DHS/1) to record the latitude, longitude, and altitude for the cluster and segment on paper. First, the mapper will write down the latitude and longitude coordinates in decimal degree format and altitude in meters in the Identification Box on the "Location Map Cluster" Form (DHS/1). Second, the mapper will draw a *circle,* in the middle of the cluster/segment, at the location where he/she captured the waypoint.

After the listing is complete, the GPS units must be collected as soon as possible and returned to the sampling office by the Mapping and Listing Coordinator. The waypoints will then be downloaded and examined for problems by the designated sampling staff. The Sampling Coordinator should

[6] This number is different from the household number later given to all of the households listed in the whole cluster just prior to household selection.

designate one member of the Data Processing Team to receive and process the GPS waypoint file and then give the file to survey manager.

In most situations, the Mapping and Listing Coordinator will be responsible for providing the listing teams with a GPS unit prior to the listing. Before these units are distributed they should be set up for use by the listers. For DHS surveys, the only format which is acceptable is Decimal Degrees, regardless of what geographic standards may be in use for other purposes. To set the format, enter the SETUP menu and in the UNITS sub-menu, select the item POSITION FRMT and press the ENTER button. Select "hddd.ddddd" Decimal Degrees, which is the first item. Once "hddd.ddddd" is highlighted, press the ENTER button. It is important that all the GPS units be set up in the same way so that the waypoints returned at the end of the survey are all in the same format. For more details on how to properly prepare the GPS units for waypoint collection, please refer to the DHS *Manual for GPS Data Collection.*

2.7 Listing of households

The lister will use the Household Listing Form (Form DHS/2) to record all households found in the cluster. Begin by entering the identification information for the cluster. The first two columns are reserved for office use only—leave them blank.

Complete the rest of the form as follows:

Column (1) [*Serial Number of Structure*]: For each structure, record the same structure serial number that the mapper enters on the sketch map. All the structures recorded on the sketch map (except the landmarks) must be recorded on the listing form and numbered.

Column (2) [*Address/description of Structure*]: Record the street address of the structure. Where structures do not have visible street addresses (especially in rural areas), give a description of the structure and any details that help in locating it (for example, in front of the school, next to the store, etc.).

Column (3) [*Residence Y/N*]: Indicate whether the structure is used for residential purposes (eating and sleeping) by writing Y for "Yes". In cases where a structure is used for commercial or other purposes, write N for "No". Structures used both for residential and commercial purposes (for example, a combination of store and home) should be classified as residential (i.e. mark Y in column 3). Make sure to list any household unit found in a nonresidential structure (for example, a guard living inside a factory or in a church). Also do not forget to list vacant structures and structures under construction, and in Column (6) give some explanation (for example: vacant, under construction, etc.) All structures seen in the cluster should be recorded on the sketch map of the cluster and in the listing.

Column (4) [*Serial Number of Household in Structure*]: This is the serial number assigned to each household found in the structure; there can be more than one household in a structure. The first household in the structure will always have number "1". If there is a second household in the structure, then this household should be recorded on the next line, a "2" is recorded in Column (4), and Columns (1) to (3) repeat the structure number and address or are left blank.

Column (5) [*Name of Head of Household*]: Write the name of the head of the household. There can only be one head per household. If no one is home or the household refuses to cooperate, ask neighbors for the name of the head of the household. If a name cannot be determined, leave this column blank. Note that it is not the name of the landlord or owner of the structure that is needed, but the name of the head of the household that lives there.

Column (6) [*Observations/Occupied or not*]: This space is provided for any special remarks that might help the coordinator decide whether to include a household in the household selection or not, and might also help the interviewing team locate the structure or identify the household during the main survey fieldwork.

If the structure is an apartment block or block of flats, assign one serial number to the entire structure (only one square with one number appears on the sketch map), but complete Columns (2) through (6) for each apartment in the structure individually. Each apartment should have its own address, which is the apartment number within the structure.

The listing team should be careful to locate hidden structures. In some areas, structures may have been built so haphazardly that they are easily missed. In rural areas, structures may be hidden by tall grasses and trees. If there is a pathway leading from the listed structure, check to see if the pathway goes to another structure. Talking with people living in the area may help in identifying the hidden structures.

2.8 Segmentation of large clusters

A certain number of the selected EAs may be very large in population size. A complete listing of EAs that are very large may not be feasible for the survey. These EAs should be subdivided into several smaller segments, only one of which will be included in the survey and listed. In this case, the DHS cluster corresponds to a segment of an EA. When the team arrives in a large EA that may need segmentation, it should first tour the EA and make a quick count to get the estimated number of households residing in the EA. There is no standard threshold for the size of an EA that needs to be segmented, or for segment size. But for efficiency and accuracy considerations, DHS recommends that if the EA size is bigger than 300 households, then the team should communicate to the coordinator the cluster number, the estimated number of households and the suggested number of segments to be created. The final decision to segment an EA, and the number of segments to be created, can only be taken by the coordinator. Ideally, for ease of operation, an EA would only need to be segments into 2 segments, with an ideal segment size of 150-200 households in each segment. Dividing an EA into a large number of segments (more than 3) should be avoided if it is not really necessary in order to minimize errors.

In dividing an EA into segments, the ideal would be to have segments of approximately equal size, but it is also important to adopt segment boundaries that are easily identifiable. In the first tour of the cluster draw a location map of the entire cluster. Using identifiable boundaries such as roads, streams, and electric power lines, divide the EA into the designated number of roughly equal-sized segments. On the location map of the EA, show clearly the boundaries of the segments created. Number the segments sequentially. Estimate the relative size of each segment in the following manner: quickly count the number of dwellings in each segment, add up the total number of dwellings in the EA and calculate the proportion of the dwellings in the whole EA that are located in each segment.

Example 2.1: A cluster of 620 dwellings has been divided into 3 segments and the results are as follows:

Segment 1:	220 dwellings,	or	220/620	=	35 percent
Segment 2:	190 dwellings,	or	190/620	=	31 percent
Segment 3:	210 dwellings,	or	210/620	=	34 percent
Total:	620 dwellings,	or	620/620	=	100 percent

On Form DHS/3 (Segmentation Form) write the size of the segments in the appropriate columns (number and percent) and calculate the cumulative size of all of the segments in terms of a percentage. The cumulative size of the last segment on the list must be equal to 100.

Segment number	Number of dwellings	Percent	Cumulative percent
1	220	35	35
2	190	31	66
3	210	34	100

For each large EA to be segmented, a random number between 0 and 100 will be selected in the central office and included in the file. Compare this random number with the cumulative size. Select the first segment for which the cumulative size is greater than or equal to the random number.

Random number: 67
Segment selected: Segment number 3

Proceed with the household listing operation in segment number 3 as described in the above sections (see Appendix 2.3 for an example of how to complete the segmentation form.) Draw a detailed sketch map of the selected segment and list all the households found in the selected segment.

2.9 Quality control

To ensure that the work done by each listing team is acceptable, quality checks should be performed. The coordinator should tour the regions during the household listing operation, and assess the quality of the finished clusters. The coordinator should select a finished cluster and do an independent listing of 10 percent of the cluster. If important errors are found, the whole cluster should be relisted. If the problem is related to systematic errors, and it is not possible to do corrections on the listing forms, then all of the listed clusters should be relisted.

2.10 Prepare the household listing forms for household selection

Once the central office receives the completed listing materials for a cluster, they must first assign a serial number to all of the households in the cluster in the second column of the form DHS/2. Only occupied residential households (including households that refused to cooperate at the time of listing and households where the occupants were absent at the time of listing but would return shortly and would be at home during the period of household interview) will be numbered. This is a continuous serial number from 1 to the total number of occupied residential households listed in the cluster. Leave the cell in the second column blank if the household is not occupied, or if the structure is not a residential structure. Fill in the second column only if the structure on that row is an occupied household. Make sure that the numbering of all occupied households follows sequentially from the previous occupied household on the list, with no gaps or repetitions in the numbering. See the example of a completed listing form in Appendix 2.3.

After assigning the serial numbers to all households listed in the cluster, copy the total number of households listed to the column “Number of households listed” in the Excel file prepared for household selection. Make sure this number is recorded in the correct row for the cluster number. In the column “Segmentation information” record the percentage of the entire EA population that is included in the selected segment. The segmentation information is important for correctly calculating

the sampling weights. After the total number of households listed in the cluster has been entered in the Excel file, the spreadsheet automatically generate the household numbers of those households selected to be interviewed. Copy the numbers of the selected households to the first column of the form DHS/2, corresponding to the serial number of the households in the listing form. These are the households that must be interviewed. It is recommended to use a different colored pen on the listing forms to indicate the households selected for interviewing. It is also very helpful to use color on the cluster's sketch map to mark the structures where the selected households are located.

In many surveys, a sub-sample of households will be selected for the men's survey. The household selection spreadsheet uses shaded columns to indicate which households are selected for the men's survey. Put a mark in the first column on the form DHS/2 next to the number of the selected household to indicate the households selected for the men's survey, or use a different colored pen for the households selected for both men's and women's surveys. Make a copy of the whole package of files (sketch maps and the listing forms with household selection). Give the original to the interviewing team for the household interview and keep the other copy in the central office.

Appendix 2.1 Example listing forms

Form DHS/1 **PAGE 1 of 3**

Map Information Form

Identification Label	***Code***
Locality _____ DHS Cluster Number..... Urban/Rural (Urban=1/Rural=2) ..... EA Number ..... District _____ Region _____ Name of Mapper _____ Name of Lister _____	

GPS Unit Tracking Number ..... Waypoint name (entered in GPS unit) ..... Latitude (North/South) ..... N / S Longitude (East/West) ..... E / W Altitude / Elevation (Meters)	

Observations:

Road access _____

_____

_____

Other useful information _____

_____

_____

Locality _____ **District** _____

Location map **DHS Cluster:** | | | |

Locality _____ **District** _____

Sketch map of cluster **DHS Cluster:** | | | |

Form DHS / 2

Page |__|__| of |__|__| pages

Form of cartography

DHS CLUSTER: |___|___|___| Urban-Rural : .....|___| Village : ..... |__|__|

Province :.....|__|__|	Commune : ..... |__|__|	Mapper : ..... |__|__|
District :..... : |__|__|	EA Number :.....|__|__|	Lister :..... |__|__|

Leave blank		Serial number of structure (1)	Address / Description of structure (2)	Residence Y/N (3)	Serial number of household in the structure (4)	Name of household head (5)	Observation/Occupancy status (6)
Male Survey	Serial number of household						

Form DHS/3 ***Segmentation Form***

Identification Label	***Code***
Locality _____ DHS Cluster Number..... Urban/Rural (Urban=1/Rural=2) ..... EA Number ..... District _____ Region _____ Name of Mapper _____ Name of Lister _____	

Number of segments: ☐

Segment number	Number of households	Percent	Cumulative percent
1			
2			
3			
4			
5			

Random number: _____

Segment selected: _____

Appendix 2.2 Symbols for mapping and listing

Orientation to the North

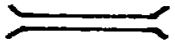
Boundaries of the cluster

Paved road

Unpaved (dirt) road

Footpath

River, creek, etc.

Bridge 

Lake, pond, etc.

Mountains, hills

Water point (wells, fountain, etc.)

Market

School

Administrative structure

Church, temple

Mosque

Cemetery

Residential structure

Non-residential structure

Vacant structure

Hospital, clinic, etc.

Electric pole

Tree or bush

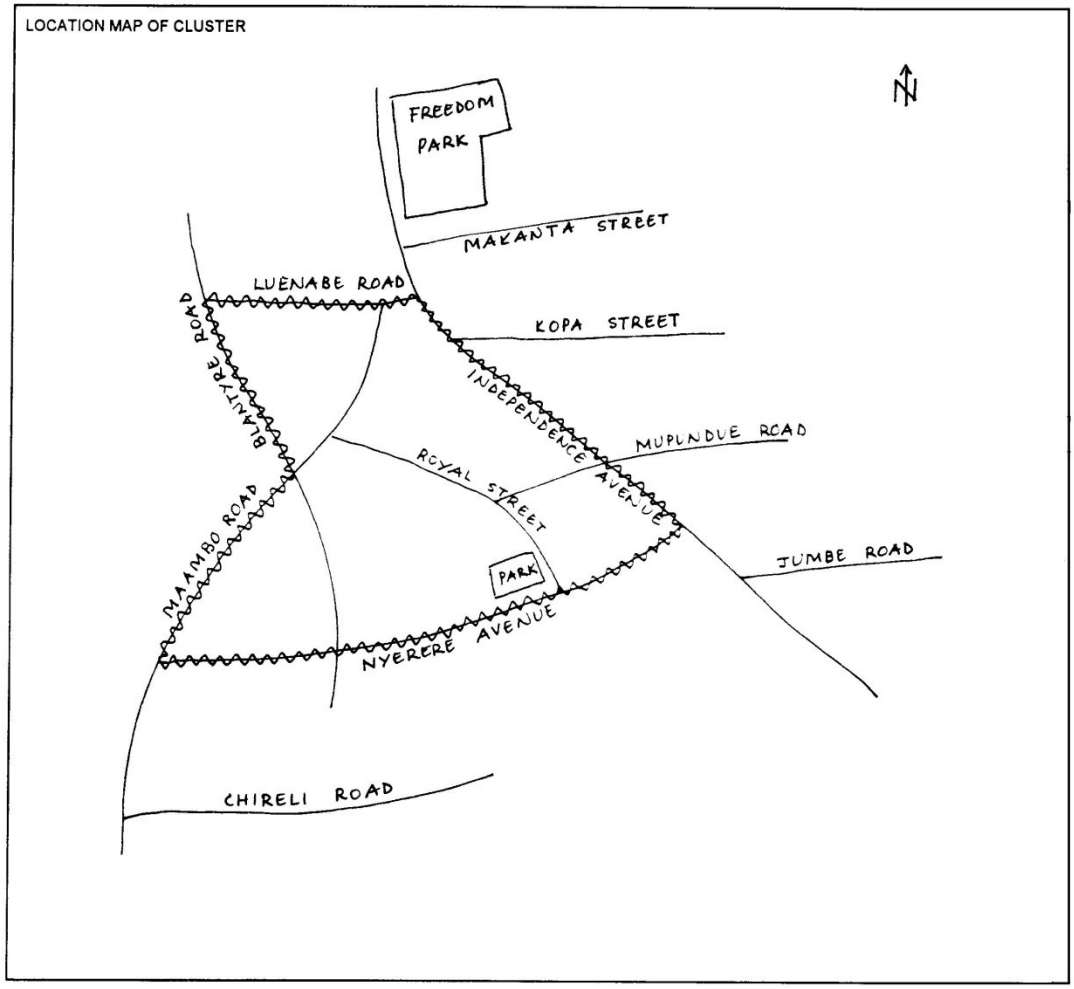
Appendix 2.3 Examples of completed mapping and listing forms

Form DHS/1 | DEMOGRAPHIC AND HEALTH SURVEY MAP INFORMATION | Page 1

IDENTIFICATION			
PROVINCE	KAYES	PROVINCE CODE	1
DISTRICT	DIEMA	DISTRICT CODE	04
TOWN/VILLAGE	DIEMA	TOWN/VILLAGE CODE	02
NAME OF MAPPER	Harrison Sidibe	CLUSTER CODE	017
NAME OF LISTER	John Melaku	DHS CLUSTER N°	001

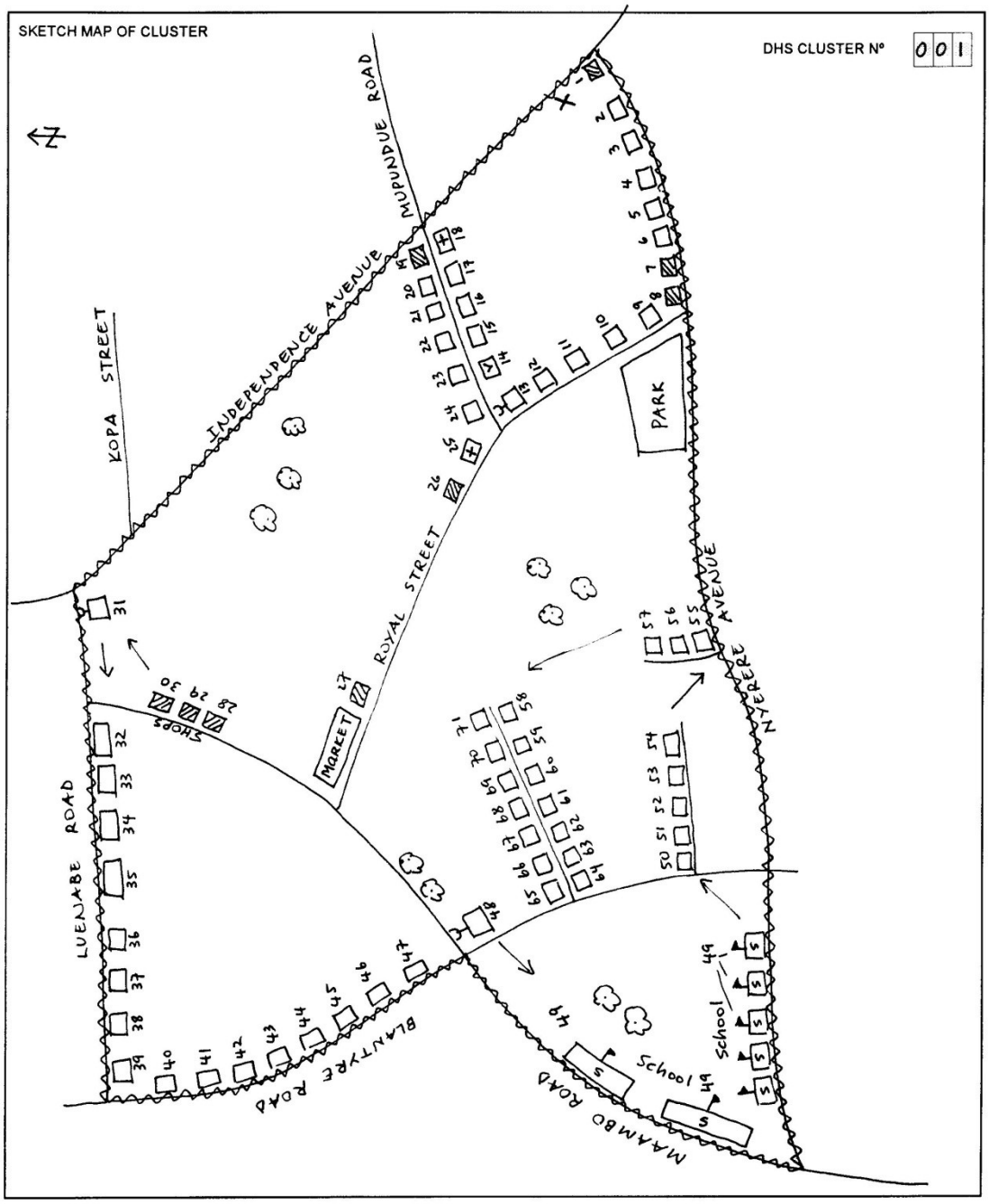
OBSERVATIONS:

LOCATION MAP OF CLUSTER



SKETCH MAP OF CLUSTER

DHS CLUSTER Nº 0 0 1



Form DHS/2

DEMOGRAPHIC AND HEALTH SURVEY
HOUSEHOLD LISTING FORM

Page 1 of 7 pages

DHS CLUSTER N° 001

LEAVE BLANK		SERIAL N° OF STRUC-TURE (1)	ADDRESS/DESCRIPTION OF STRUCTURE (2)	RESIDENCE Y/N (3)	SERIAL N° OF HOUSEHOLD IN STRUCTURE (4)	NAME OF HEAD OF HOUSEHOLD (5)	OBSERVATIONS (6)
HH TO INTERVIEW	HH NUMBER						
		1	Nyerere Avenue	N			Pharmacy Star
	1	2	6 Nyerere Avenue	Y	1	Diane Obote	
	2	3	8 Nyerere Avenue	Y	1	Eugene Kariba	
	3				2	Dorothy Uchi	
		4	10 Nyerere Avenue	Y	1		No one at home.
	4	5	12 Nyerere Avenue	Y	1	Sam Lowa	
	5	6	14 Nyerere Avenue	Y	1	Harrison Coulibali	
	6				2	Paul Liande	
	7				3	Harry Fiwale	
		7	Avenue Nyerere	N			In construction
		8	Nyerere Avenue	N			In construction
	8	9	22 Royal Street	Y	1	George Sidibi	
	9	10	20 Royal Street	Y	1		Refused
	10	11	18 Royal Street	Y	1	Chief Seidou	
	11	12	16 Royal Street	Y	1	Ann Tonde	
		13	Mupundue Road	N			Mosque
		14	4 Mupundue Road	N			Vacant
	12	15	6 Mupundue Road	Y	1	Suzanne Ibenga	
	13	16	8 Mupundeee Road	Y	1	David Chouta	
	14	.			2	Joseph Lupiya	
	15	17	10 Mupundue Road	Y	1	Eleni Fahmi	
	16	18	10A Mupundue Road	Y	1	Doctor Tadesse	Home upstair, clinic downstairs
	17	19	12 Mupundue Road	Y	1	Sam Sidibi	

Form DHS/3

DEMOGRAPHIC AND HEALTH SURVEY
SEGMENTATION FORM

IDENTIFICATION		
PROVINCE KOULIKORO	PROVINCE CODE	4
DISTRICT DIOLA	DISTRICT CODE	02
TOWN/VILLAGE DIONGAGA	TOWN/VILLAGE CODE	06
NAME OF MAPPER WOLDE CONATE	CLUSTER CODE	023
NAME OF LISTER ANDRE LUENA	DHS CLUSTER Nº	015

NUMBER OF SEGMENTS TO BE CREATED 03

Segment Number	Number of dwellings	Percent	Cumulative percent
1	220	35%	35%
2	190	31%	66%
3	210	34%	100%
4			
5			

RANDOM NUMBER BETWEEN 1 AND 100: 067

SEGMENT SELECTED: 03

3 SELECTED SAMPLING TECHNIQUES

In this section, some of the most commonly used sampling techniques and their application are presented. The presentation will focus mainly on practical rather than theoretical aspects. However, the chapter does touch on some basic theoretical properties of the techniques used in the DHS surveys.

We focus on without replacement sampling rather than with replacement sampling procedures, since the latter represents a reduction of efficiency for samples of a fixed size due to the potential that some sampling units may be repeated. When this occurs, the amount of information carried in a fixed size sample is reduced because the same sampling unit is selected several times. For readers who are interested in the theoretical aspects of the selected sampling techniques, please refer to the textbooks dealing with survey sampling theory listed in the references.

3.1 Simple random sampling

We begin with *simple random sampling without replacement* (*SRSWOR*) since this is a fundamental sampling procedure that is used as standard to which the efficiency of other sampling procedures is compared. Simple random sampling without replacement is a selection procedure where every unit has an equal chance of being selected. Selection can be performed through successive draws without replacement from a well-mixed container containing all sampling units, or using certain computerized algorithms to select from a list of all sampling units.

Let *N* be the total number of sampling units, let n be the total sample size, n<N. The probability of selection for every i^{th} unit is given by:

$$P_i = \frac{n}{N}$$

The design weight (assuming no non-response) is given by:

$$D_i = 1/P_i = \frac{N}{n}$$

The probability for any particular *n* different units selected together in a sample *s* is given by:

$$P_s = 1/\binom{N}{n}$$

where $\binom{N}{n}$ is the total number of combinations of *n* elements out of *N*. Let $y_1, y_2, \ldots y_n$ be the observations made from the selected units on a variable of interest, then the weighted sample mean which is the same as the unweighted sample mean,

$$\bar{y} = \sum_1^n D_i y_i / \sum_1^n D_i = \frac{1}{n}\sum_1^n y_i$$

,

is an unbiased estimator of the population mean, $\bar{Y} = \frac{1}{N}\sum_1^N y_i$, with its sampling variance given by

$$V_{srs}(\bar{y}) = \frac{1-f}{n} S_y^2$$

where $\sigma_y^2 = \frac{1}{N-1}\sum_1 (y_i - \bar{y})^2$ is the finite population variance of the variable y and f=*n/N* is the sampling fraction. An unbiased estimation of this variance can be made using

$$v_{srs}(\bar{y}) = \frac{1-f}{n} s_y^2$$

where $s_y^2 = \frac{1}{n-1}\sum_1 (y_i - \bar{y})^2$ is the sample variance. When *n* and *N* are large, the standardized variable

$$\frac{\bar{y} - \bar{Y}}{SE(\bar{y})}$$

follows a *student-t* distribution with n-1 degrees of freedom and $SE(\bar{y})$ is the square root of $v_{srs}(\bar{y})$. Therefore the confidence limits of the population mean $\bar{Y}$ can be constructed based on sample observations allowing for 95% confidence that the true value of $\bar{Y}$ will lie within the range of $\bar{y} - 1.96 * SE(\bar{y})$ and $\bar{y} + 1.96 * SE(\bar{y})$. DHS reports use $\bar{y} \pm 2 * SE(\bar{y})$ for a conservative estimate of 95% confidence limits.

Given a complete list of all sampling units in a computerized file, the easiest way to draw a simple random sample of size *n* is to first generate a uniformly distributed random number between 0 and 1 and associate a number with each of the sampling units. Next, sort the file based on the generated random numbers in ascending order, and the first *n* units associated with the *n* smallest random numbers are the selected units. This procedure provides a *SRSWOR* sample of size *n*. This procedure is easy to implement, but requires sorting of the sampling frame. Since sorting is time consuming, the following algorithm (Tillé, 2001) may be used with the sampling frame without sorting:

Definition of terms and the initial step	
k: the k^{th} unit of the frame file; *j*: the j^{th} selected unit	
k = 0	
j = 0	
repeat if *j* < *n*	generate a uniformly distributed random number between [0,1) if *u* < $\frac{n-j}{N-k}$ then unit *k* + 1 is selected; *j* = *j* + 1 else unit *k* + 1 is not selected *k* = *k* + 1

3.2 Equal probability systematic sampling

3.2.1 Sampling theory

Systematic sampling (*SYS*) is the selection of sampling units at a fixed interval from a list, starting from a randomly determined point. Selection is systematic because selection of the first sampling unit determines the selection of the remaining sampling units. Compared with *SRSWOR,* systematic sampling has the following advantages:

1) It is easier to perform;

2) It allows easy verification of the selection;

3) If the sampling frame is in some order, it provides a stratification effect with respect to the variables on which the frame is sorted, and with a proportional allocation. This stratification is called *implicit stratification*.

4) Implicit stratification prevents unexpected concentration of sample points in certain areas such as is possible with *SRSWOR*.

Because of these advantages, especially (3) and (4), systematic selection is more often used than simple random sampling.

Systematic sampling is normally carried out as follows: assuming a whole number interval $I=N/n$, where N is the number of units in the frame list and n is the number of units to be selected. The procedure begins with an integer random number S that is less than or equal to I. The units to be selected are S, S+I, S+2*I, ..., S+(n-1)*I. When I is not a whole number there may be appreciable errors in rounding it to the nearest whole number, it is suggested that the decimal interval method be used. Selection with a decimal interval may be carried out as follows:

1) Calculate the interval I rounded to two decimal places.

2) Generate a random number R between 0 and *1* with two decimal points.

3) Compute the sequence of sampling numbers: $R*I$, $R*I + I$, $R*I$ + 2*I, ..., $R*I + (n - 1)*I$

4) Round up the above calculated sampling numbers to the next highest whole numbers; these are the selected units' numbers.

Example 3.2.1:

Let N=100, n=14, so that I=7.14; let the generated random number be R=0.96. The sampling numbers and the corresponding selected unit numbers are as follows:

6.85	13.99	21.13	28.27	35.41	42.55	49.69	56.83	63.97	71.11	78.25	85.39	92.53	99.67
7	14	22	29	36	43	50	57	64	72	79	86	93	100

In this example, the decimal interval method gives a selection interval which is sometimes 7 or sometimes 8. The household selection templates are all programmed with decimal sampling intervals.

Often sample design requires numerous systematic samples as is the case when a systematic sample of households is needed within each selected cluster. In this situation a separate random start R should be determined independently for each cluster.

With *SYS*, the probability of selection for any unit i is given by

$$P_i=\frac{1}{I}=\frac{n}{N}$$

The design weight (assuming no non-response) is given by

$$D_i=1/P_i=\frac{N}{n}$$

Let $y_1, y_2, \ldots y_n$ be the observations made from the selected units on a variable of interest, then the weighted sample mean which is the same as the unweighted sample mean

$$\bar{y}=\sum_1^n D_i y_i / \sum_1^n D_i = \frac{1}{n}\sum_1^n y_i$$

is an unbiased estimator of the population mean $\bar{Y}=\frac{1}{N}\sum_1^N y_i$. For simplicity, assuming an integer sampling interval *I*, the sampling variance of the sample mean is given by

$$V_{sys}(\bar{y})=\frac{(1-1/N)}{n}S_y^2\left[1+(n-1)\rho_w\right]$$

where $S_y^2=\frac{1}{N-1}\sum_1^N\left(y_k-\bar{Y}\right)^2$ is the population variance; ρ_w is the correlation coefficient between pairs of units in the same systematic sample. When ρ_w is negative, *SYS* is more precise than *SRSWOR*; when ρ_w is positive, *SYS* is less precise than *SRSWOR*. Unlike the case of *SRSWOR*, the variance estimate

$$\upsilon_{sys}(\bar{y})=\frac{1-f}{n}s_y^2$$

is not an unbiased estimate of the sampling variance; where $s_y^2=\frac{1}{n-1}\sum_1^n\left(y_i-\bar{y}\right)^2$ is the sample variance. However, $\upsilon_{sys}(\bar{y})$ is the special case of the recommended Hartley-Rao (1962) estimator in the case of un-equal probability systematic sampling. $\upsilon_{sys}(\bar{y})$ is equivalent to treating the systematic sample as if it was drawn by *SRSWOR*, and therefore is called an "estimator with simple random sampling approximation".

Theoretically, with SYS there is no unbiased estimator for the variance of the sample mean since systematic sampling is equivalent to randomly selecting one sample among the *I* possible samples. This is a major drawback for the *SYS*. However, when the sampling units in the frame file do not present any linear trend in the variable of interest, nor periodic changes, or the units are randomly ordered, $\upsilon_{sys}(\bar{y})$ is a good approximation of the sampling variance $V_{sys}(\bar{y})$. When there is a linear trend in the variable of interest, assuming the selection of the k^{th} systematic sample, where the summation is over non-overlapping successive units, the following estimator (Wolter, 1984; Wolter 1985) is a better approximation of $V_{sys}(\bar{y})$:

$$\upsilon_{sys}^{i}(\bar{y})=\frac{1-f}{n}\frac{1}{n}\sum_1^{\lfloor n/2\rfloor}\left(y_{k+(j-1)*I}-y_{k+j*I}\right)^2$$

However, when confidence limits are required, $\upsilon_{sys}(\bar{y})$ is preferred because of its high coverage rates of the true population mean. It should be noted that the properties of $\upsilon_{sys}^{i}(\bar{y})$ are different from the collapsed strata estimator for stratified sampling with one unit per stratum because the successive observations in a *SYS* sample are probability-one correlated, while the collapsed strata estimator for stratified sampling has a set of completely independent observations.

When *n* and *N* are large, the sample mean has the same asymptotic properties as that of the simple random sample mean; therefore confidence intervals can be constructed in a similar way to those for a simple random sample.

3.2.2 Excel templates for systematic sampling

The MEASURE DHS program has developed Excel templates that can be used for equal probability systematic sampling of households. The templates can be used to perform simple selection, selection with runs, self-weighting selection without sample size control and self-weighting selection with sample size control. Figure 3.1 below shows a portion of the simple selection procedure with a sample take of 20 households per cluster. The darker shaded areas require data input. The area to the left of the column labeled, “Num HH listed” is reserved for cluster IDs. Numbers for the selected households are shown to the right of the column labeled “Random (0-1)”. Figure 3.2 below shows a portion of the selection procedure with runs of 4 households. Both selections incorporate a selection of a sub-sample. Figure 3.3 shows a simple self-weighting selection with an average sample take of 20 households, without sample size control, but with the minimum and maximum number of sample takes of 10 and 30 households respectively.

Figure 3.4 shows a self-weighting selection, with runs, with an average sample take of 20 households per cluster, without sample size control, but with minimum and maximum sample takes of 10 and 30 households respectively; both of the selections incorporate a sub-sample of 10 households per cluster. Note that the selection procedure with runs is circular, meaning that when the selection interval is not an integer, and when the run is not a divisor of the total number of households listed, then the last selected household number may be smaller than the first selected household number.

Figures 3.5 and 3.6 show self-weighting selections with sample size control; the control area is the sampling stratum. The disadvantage of the self-weighting selection with sample size control is that the selection procedure will do the household selection only if the household listing results are entered for the entire control area. This condition may represent a constraint in some situations.

Figure 3.7 shows a manual selection carried out in the field that can be performed easily using a simple calculator. If household selection at the central office is not feasible; the interviewer can perform the household selection in the field. The numbers in red represent information that is entered and the calculated terms. This procedure requires a traditional household listing operation where households are numbered and listed on household listing forms. Using the total number of households listed and the number of households to be selected, the interviewer can first calculate the selection interval then use the random number, R, associated with the selected cluster, to calculate the first sampling number or term t_1 and enter the first term to the cell for t_1. For the subsequent sampling numbers or terms, the interviewer adds the sampling interval to the previous sampling number or term. After the calculation of the sampling numbers, the interviewer should round the sampling numbers to integers in the next column; these are the selected household numbers. The interviewer is asked to copy the address and the name of the head of household of the selected households from the household listing form. The household selection form is subject to review by the field work supervisor.

Figure 3.1 Simple household selection with a sub-sample

						HOUSEHOLD SELECTION																		
				Run size	**1**																			
		Sub-sample take per cluster			**10**	1		2		3		4		5		6		7		8		9		10	
Cluster Num		Num HHs Listed	Num Selected	Select interval	Random (0-1)	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
1		**138**	**20**	6.90	0.03800	1	8	15	21	28	35	42	49	56	63	70	77	84	90	97	104	111	118	125	132
2		**151**	**20**	7.55	0.65268	5	13	21	28	36	43	51	58	66	73	81	88	96	104	111	119	126	134	141	149
3		**182**	**20**	9.10	0.97489	9	18	28	37	46	55	64	73	82	91	100	109	119	128	137	146	155	164	173	182
4		**129**	**20**	6.45	0.41931	3	10	16	23	29	35	42	48	55	61	68	74	81	87	94	100	106	113	119	126
5		**180**	**20**	9.00	0.53756	5	14	23	32	41	50	59	68	77	86	95	104	113	122	131	140	149	158	167	176
6		**173**	**20**	8.65	0.70405	7	15	24	33	41	50	58	67	76	84	93	102	110	119	128	136	145	154	162	171
7	**C**	**140**	**20**	7.00	0.51868	4	11	18	25	32	39	46	53	60	67	74	81	88	95	102	109	116	123	130	137
8	**l**	**69**	**20**	3.45	0.25579	1	5	8	12	15	19	22	26	29	32	36	39	43	46	50	53	57	60	63	67
9	**u**	**176**	**20**	8.80	0.96775	9	18	27	35	44	53	62	71	79	88	97	106	115	123	132	141	150	159	167	176
10	**s**	**90**	**20**	4.50	0.40192	2	7	11	16	20	25	29	34	38	43	47	52	56	61	65	70	74	79	83	88
11	**t**	**131**	**20**	6.55	0.32702	3	9	16	22	29	35	42	48	55	62	68	75	81	88	94	101	107	114	121	127
12	**e**	**92**	**20**	4.60	0.76363	4	9	13	18	22	27	32	36	41	45	50	55	59	64	68	73	78	82	87	91
13	**r**	**126**	**20**	6.30	0.41681	3	9	16	22	28	35	41	47	54	60	66	72	79	85	91	98	104	110	117	123
14		**199**	**20**	9.95	0.84599	9	19	29	39	49	59	69	79	89	98	108	118	128	138	148	158	168	178	188	198
15		**225**	**20**	11.25	0.91906	11	22	33	45	56	67	78	90	101	112	123	135	146	157	168	180	191	202	213	225
16	**I**	**205**	**20**	10.25	0.12089	2	12	22	32	43	53	63	73	84	94	104	114	125	135	145	155	166	176	186	196
17	**D**	**148**	**20**	7.40	0.88941	7	14	22	29	37	44	51	59	66	74	81	88	96	103	111	118	125	133	140	148
18		**146**	**20**	7.30	0.25095	2	10	17	24	32	39	46	53	61	68	75	83	90	97	105	112	119	126	134	141
19		**139**	**20**	6.95	0.14534	2	8	15	22	29	36	43	50	57	64	71	78	85	92	99	106	113	120	127	134
20		**201**	**20**	10.05	0.84179	9	19	29	39	49	59	69	79	89	99	109	120	130	140	150	160	170	180	190	200

2

Figure 3.2 Selection of runs with a sub-sample

						HOUSEHOLD SELECTION																		
				Run size	4																			
		Sub-sample take per cluster			**10**		1		2		3		4		5		6		7		8		9	10
Cluster Num		Num HHs Listed	Num Selected	Select interval	Random (0-1)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1		**138**	**20**	6.90	0.77576	21	22	23	24	49	50	51	52	77	78	79	80	105	106	107	108	129	130	131	132
2		**151**	**20**	7.55	0.05693	1	2	3	4	29	30	31	32	61	62	63	64	93	94	95	96	121	122	123	124
3		**182**	**20**	9.10	0.10590	1	2	3	4	41	42	43	44	77	78	79	80	113	114	115	116	149	150	151	152
4		**129**	**20**	6.45	0.64741	17	18	19	20	41	42	43	44	69	70	71	72	93	94	95	96	117	118	119	120
5		**180**	**20**	9.00	0.60810	21	22	23	24	57	58	59	60	93	94	95	96	129	130	131	132	165	166	167	168
6		**173**	**20**	8.65	0.96364	33	34	35	36	65	66	67	68	101	102	103	104	137	138	139	140	169	170	171	172
7	**C**	**140**	**20**	7.00	0.11160	1	2	3	4	29	30	31	32	57	58	59	60	85	86	87	88	113	114	115	116
8	**l**	**69**	**20**	3.45	0.15540	1	2	3	4	13	14	15	16	29	30	31	32	41	42	43	44	57	58	59	60
9	**u**	**176**	**20**	8.80	0.00870	1	2	3	4	33	34	35	36	69	70	71	72	105	106	107	108	141	142	143	144
10	**s**	**90**	**20**	4.50	0.32205	5	6	7	8	21	22	23	24	41	42	43	44	57	58	59	60	77	78	79	80
11	**t**	**131**	**20**	6.55	0.69849	17	18	19	20	45	46	47	48	69	70	71	72	97	98	99	100	121	122	123	124
12	**e**	**92**	**20**	4.60	0.51119	9	10	11	12	25	26	27	28	45	46	47	48	65	66	67	68	81	82	83	84
13	**r**	**126**	**20**	6.30	0.31826	9	10	11	12	33	34	35	36	57	58	59	60	81	82	83	84	109	110	111	112
14		**199**	**20**	9.95	0.69129	25	26	27	28	65	66	67	68	105	106	107	108	145	146	147	148	185	186	187	188
15		**225**	**20**	11.25	0.67523	29	30	31	32	73	74	75	76	121	122	123	124	165	166	167	168	209	210	211	212
16	**I**	**205**	**20**	10.25	0.30267	13	14	15	16	53	54	55	56	93	94	95	96	133	134	135	136	177	178	179	180
17	**D**	**148**	**20**	7.40	0.53373	13	14	15	16	45	46	47	48	73	74	75	76	105	106	107	108	133	134	135	136
18		**146**	**20**	7.30	0.32483	9	10	11	12	37	38	39	40	65	66	67	68	97	98	99	100	125	126	127	128
19		**139**	**20**	6.95	0.69275	17	18	19	20	45	46	47	48	73	74	75	76	101	102	103	104	129	130	131	132
20		**201**	**20**	10.05	0.34623	13	14	15	16	53	54	55	56	93	94	95	96	133	134	135	136	173	174	175	176

9

Figure 3.3 Simple self-weighting selection without sample size control

| | | | | | | | | | Household selection | | | | | | | | |

| | | |

Figure 3.4 Self-weighting selection with runs and without sample size control

| | | | | | | | | | Household selection | | | | | | | | |

	0.00823						0.5775		
18	9	63	0.002615		90	29	3.10	1	6 7 8 9 10 21 22 23 24 25 36 37 38 39 40 56 57 58 59 60 71 72 73 74 75 86 87 88 89
	0.01660						0.8980		
19	8	127	0.002615		98	15	6.53	1	26 27 28 29 30 61 62 63 64 65 91 92 93 94 95
	0.00941						0.7108		
20	6	72	0.002615		75	21	3.57	6	11 12 13 14 15 31 32 33 34 35 46 47 48 49 50 66 67 68 69 70 6

Figure 3.5 Self-weighting selection with sample size control

| | | | | Num of HHs expected | | Num of HHs selected | | | Household selection | | | |

| | | |

Figure 3.6 Self-weighting selection with runs and with sample size control

| | | | | Num of HHs expected | | Num of HHs selected | | | Household selection | | | |

| | | |

Figure 3.7 Manual household selection in the field

Form DHS/4			**Household Selection Form**	
Cluster ID I__I__I__I			Locality:.....	Urban-Rural : ..... I__I
EA number: I__I__I__I			Region : .....I__I__I	District : .....I__I__I
Number of households listed N= 126			Number of households to be selected n= 25	
Random number (0, 1) R = 0.43			Selection interval I= N/n = 5.04	Term t_1 = I×R = 2.17
SN° of selection k	Term (k>1) $t_k = t_{k-1} + I$	Selected HH number	Address of household	Name of household head
1	2.17	3		
2	7.21	8		
3	12.25	13		
4	17.29	18		
5	22.33	23		
6	27.37	28		
7	32.41	33		
8	37.45	38		
9	42.49	43		
10	47.53	48		
11	52.57	53		
12	57.61	58		
13	62.65	63		
14	67.69	68		
15	72.73	73		
16	77.77	78		
17	82.81	83		
18	87.85	88		
19	92.89	93		
20	97.93	98		
21	102.97	103		
22	108.01	109		
23	113.05	114		
24	118.09	119		
25	123.13	124		
26				
27				
28				
29				
30				
Notes: Random number R is between 0 and 1, with two decimal places				

3.3 Probability proportional to size sampling

3.3.1 Sampling theory

In order to increase sampling efficiency, a sampling procedure can attribute different selection probabilities to different sampling units. In general, a "large" sampling unit will contribute more to the sampling variance if equal probability selection is used. If large sampling units are selected with larger chances, sampling variance may be greatly reduced. To the extreme, a good strategy is to select very large sampling units with certainty or with a probability of one. Assuming that each sampling unit has some kind of known measure of size which is positively correlated with the variable of interest, a *Probability Proportional to the measure of Size* (*PPS*) selection has the same four advantages as *SYS* sampling. This procedure assigns each sampling unit a specific chance to be selected in the sample before the sampling begins, and the chance is proportional to its measure of size.

Let M_i be the measure of size of unit *i*; let $\sum_1^N M_i$ be the total measure of size; let *n* be the design sample size. A *PPS* sampling procedure will select unit *i* with a probability π_i such that

$$\pi_i = \frac{nM_i}{\sum M_i}$$

The design weight (assuming no non-response) is given by

$$D_i = 1/\pi_i = \frac{\sum M_i}{nM_i}$$

Let $y_1, y_2, \dots y_n$ be the observations made from the selected units on a variable of interest, then the weighted sum of the observations

$$\hat{y}_{PPS} = \sum_1^n D_i y_i = \sum_1^n \frac{y_i}{\pi_i}$$

is an unbiased estimator of the population total $Y = \sum_1^N y_i$. The variance of this estimator is given by

$$V(\hat{y}_{PPS}) = \frac{1}{2}\sum_{i=1}^{N}\sum_{j\neq i,\, j=1}^{N}(\pi_i\pi_j - \pi_{ij})\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2 \qquad \text{(Yates-Grundy, 1953)}$$

where π_{ij} is the joint probability of selecting units *i* and *j* together in a sample. If all the joint probabilities $\pi_{ij} > 0$, then the above variance can be estimated unbiased by:

$$\hat{V}(\hat{y}_{PPS}) = \frac{1}{2}\sum_{i=1}^{n}\sum_{j\neq i,\, j=1}^{n}\left(\frac{\pi_i\pi_j - \pi_{ij}}{\pi_{ij}}\right)\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2 \qquad \text{(Yates-Grundy, 1953)}$$

However, the above estimator is not calculable because the joint probabilities π_{ij} are usually unknown. Hartley and Rao (1962) provided an approximation of the above estimator which involves only the first order selection probabilities π_i :

$$\hat{V}_{HR}(\hat{y}_{PPS}) \simeq \frac{1}{2}\frac{1}{n-1}\sum_{i=1}^{n}\sum_{j\neq i,\, j=1}^{n}\left(1 - (\pi_i + \pi_j) + \frac{1}{n}\sum_1^N \pi_k^2\right)\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2 \qquad \text{(Hartley-Rao, 1962)}$$

But the Hartley-Rao estimator requires knowledge of the selection probability of all sampling units in the population (through $\sum_1^N \pi_k^2$) which is usually not calculated in the sample selection. The general documentation just keeps the selection probability for the selected units. By replacing $\sum_1^N \pi_k^2$ by its sample estimation $\sum_1^n \frac{\pi_i^2}{\pi_i} = \sum_1^n \pi_i$, the Hartley-Rao estimator can be further simplified (Ren, 2003)

$$\hat{V}_R(\hat{y}_{PPS}) = \frac{n}{n-1} \sum_1^n (1-\pi_i) \left(\frac{y_i}{\pi_i} - \frac{\hat{y}_{PPS}}{n} \right)^2$$

In the case of equal probability sampling, both $\hat{V}_{HR}(\hat{y}_{PPS})$ and $\hat{V}_R(\hat{y}_{PPS})$ will be reduced to the variance estimator with simple random sampling approximation. Suppose that $\pi_i < 1$ for all *i*, both Yates-Grundy and Hartley-Rao estimators may produce negative variance estimation, while $\hat{V}_R(\hat{y}_{PPS})$ is always positive.

Wolter (1984; 1985) conducted an extensive study on the variance estimation for systematic sampling, including the successive difference estimator similar to $v_{sys}^{'}(\bar{y})$. He recommends the use of the Hartley-Rao estimator if the population does not present any trends in the measure of size variable and the variable of interest, especially when a confidence interval is required.

The above results for population total estimation can be adapted to mean estimation:

$$\bar{y}_{PPS} = \sum_1^n D_i y_i / \sum_1^n D_i = \sum_1^n \frac{y_i}{\pi_i} / \sum_1^n \frac{1}{\pi_i}$$

$\bar{y}_{PPS}$ is an approximately unbiased estimator for the population mean with approximate variance given by:

$$V(\bar{y}_{PPS}) = \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N (\pi_i \pi_j - \pi_{ij}) \left(\frac{y_i - \bar{Y}}{\pi_i} \right) \left(\frac{y_j - \bar{Y}}{\pi_j} \right)$$

If the units are not specially ordered according to the variable of interest in the sampling frame, the approximate sample variance of the estimator can be estimated by

$$\hat{V}_R(\bar{y}_{PPS}) = \frac{1}{\left(\sum_1^n D_i \right)^2} \frac{n}{n-1} \sum_1^n (1-\pi_i) \left(\frac{y_i}{\pi_i} - \frac{\hat{y}_{PPS}}{n} \right)^2$$

The above estimator will be reduced to the simple random sampling approximation $v_{sys}(\bar{y})$ in case of equal probability systematic sampling.

3.3.2 Operational description and examples

There are many ways to draw a *PPS* sample, but the easiest way is the *PPS* systematic sampling summarized in the following:

1) List the sampling units with their measure of size M_i

2) Calculate the cumulative measure of size $C_k=\sum_1^k M_i$ for each unit *k*, and check that the last entry C_N equals the total measure of size $\sum_1^N M_i$

3) Let n be the number of units to be selected. Compute the sampling interval $I=\frac{\sum_1^N M_i}{n}$

4) Generate a random number *R* between 0 and *1*

5) Compute the sampling numbers *R*I*, *R*I*+*I*, *R*I*+2*I, ..., *R*I*+(*n*-1)**I*

6) For each sampling number *R*I*+(*j*-1)**I*, the j^{th} sampled unit is unit *k* if C_k is the first cumulative size bigger than the sampling number *R*I*+(*j*-1)**I*

7) Calculate the selection probability of each selected unit *j*: $\frac{n*M_j}{\sum_1^N M_i}$

The following example demonstrates how manual selection is done.

Example 3.3.1:

Let *N*=20, *n*=5, $\sum_1^{20} M_i=4004$; therefore the sampling interval $I=801$; let the generated random number be $R=305$. The sampling numbers and the selected unit numbers are as follows:

ID number	Size measure M_i	Cumulative C_k	Sampling number	j^{th} selected unit	Selection probability
1	139	139			
2	101	240			
3	184	424	305	1	0.22977
4	184	608			
5	104	712			
6	259	971			
7	219	1190	1106	2	0.273477
8	192	1382			
9	224	1606			
10	197	1803			
11	150	1953	1907	3	0.187313
12	257	2210			
13	270	2480			
14	195	2675			
15	296	2971	2707	4	0.36963
16	178	3149			
17	256	3405			
18	227	3632	3508	5	0.283467
19	247	3879			
20	125	4004			

The PPS sampling has the same advantages as equal probability systematic sampling, but with this procedure a unit may be selected more than once if the unit's measure of size is bigger than the sampling interval. These large units are said to have been selected with certainty, or are *self-*

representing units. A unit selected more than once should be segmented to form a number of smaller units corresponding to the number of times the unit is selected. The selection probabilities should be recalculated using the sizes of the segmented units. With this strategy, the total sample size is kept the same as designed and the selection probabilities of the non-certainty units do not need to be adjusted.

Another way to deal with large units consists of examining the list of units before sampling begins. Computation of the interval will reveal whether there are any units of size greater than I. The simplest solution to prevent repetition during sampling might be to split each such unit into two or more approximately equal subunits of size less than I. The split would be made first on paper only. The measure of size for the original unit is divided equally among the subunits before sampling proceeds. Later the split is "materialized," either by drawing a line on the map of the unit, or by identifying a suitable dividing line during the first field visit to the unit.

If a substantial number of the units chosen to serve as PSUs are larger than the interval I, then the choice of such units to serve as PSUs was clearly incorrect. One solution to this problem is to place all PSUs with a measure of size larger than a threshold (not necessarily greater than or equal to I) before sampling and to give them special treatment, and call them self-representing units. They are not, therefore, sampling units but strata by definition. A new type of sampling unit has to be designated to serve as PSU within these areas. For the purpose of sampling error computation, it is important to realize that the term self-representing PSU is misleading. The self-representing units are in fact strata, while the new, smaller units or sub-units within them are the true PSUs. This treatment requires re-calculating the sample allocation, and then proceeding with sample selection independently in each stratum.

An Excel template for stratified PPS or equal probability systematic sampling has been developed. Figure 3.8 below shows a portion of a blank template. Figure 3.9 shows an example of stratified PPS sampling with the strata being the urban and rural areas within each province.

Figure 3.8 Part of an Excel template for stratified sampling

Stratified systematic sampling with probability proportional to size

	Random (0, 1)								
	Stratum num								
	Stratum size								
	St Sample size								
	Stratum num								
	Stratum size								
	St Sample size								
	Stratum num								
	Stratum size								
	St Sample size								
	Stratum num								
	Stratum size								
	St Sample size								
	Col name of Dom/Region			**Col name of urban/rural**			**Col name of PSU size**		
	Total number of strata			**Total sample size**			**# of Diff PSU selected**		
Serial numb	**Dom/Region name/code**	**Urban/ rural**	**PSU Size**	**Stratum number**	**Selection Probability**	**# of times Selected**	**Stratum size**	**Stratum sample size**	**Measure size: stratum**
1									
2									
3									
4									
5									
6									
7									
8									
9									
10									

Paste the frame file below

Figure 3.9 Part of an example for a province crossed urban-rural stratified PPS sampling

Stratified systematic sampling with probability proportional to size 08/25/09

ICF Macro

by Ruilin Ren

Random (0, 1)								
Stratum num	1	2	3	4	5	6	7	8
Stratum size	15	200	15	233	19	279	35	247
St Sample size	3	10	3	17	3	20	3	17
Stratum num	9	10	11	12	13	14	15	16
Stratum size	11	151	17	215	90	292	3	116
St Sample size	2	6	4	16	5	20	2	8
Stratum num								
Stratum size								
St Sample size								
Stratum num								
Stratum size								
St Sample size								

Col name of Dom/Reg	m	Col name of urbrural	r	Col name of PSU size	q			
Total number of strata	16	Total sample size	139	# of Diff PSU selected	139			

Serial numb	Dom/Reg name/code	Urban/ rural	PSU Size	Stratum number	Selection Proba	# of times Selected	Stratum size	Stratum sam-size	Measure size-strat	Province	Code	Commune	Code	EA	HH	type
1	1	1	163	1	0.168330	0	15	3	2905	BALE	1	BOROMO	3	A001	163	1
2	1	1	250	1	0.258176	0	15	3	2905	BALE	1	BOROMO	3	A002	250	1
3	1	1	109	1	0.112565	0	15	3	2905	BALE	1	BOROMO	3	A003	109	1
4	1	1	205	1	0.211704	0	15	3	2905	BALE	1	BOROMO	3	A004	205	1
5	1	1	203	1	0.209639	0	15	3	2905	BALE	1	BOROMO	3	B005	203	1
6	1	1	155	1	0.160069	1	15	3	2905	BALE	1	BOROMO	3	B006a	155	1
7	1	1	167	1	0.172461	0	15	3	2905	BALE	1	BOROMO	3	B006b	167	1
8	1	1	170	1	0.175559	0	15	3	2905	BALE	1	BOROMO	3	B006c	170	1
9	1	1	138	1	0.142513	0	15	3	2905	BALE	1	BOROMO	3	B007	138	1
10	1	1	308	1	0.318072	0	15	3	2905	BALE	1	BOROMO	3	B008	308	1
11	1	1	240	1	0.247849	1	15	3	2905	BALE	1	BOROMO	3	B009	240	1
12	1	1	303	1	0.312909	0	15	3	2905	BALE	1	BOROMO	3	C010	303	1
13	1	1	191	1	0.197246	0	15	3	2905	BALE	1	BOROMO	3	C011	191	1
14	1	1	130	1	0.134251	0	15	3	2905	BALE	1	BOROMO	3	C012	130	1
15	1	1	173	1	0.178657	1	15	3	2905	BALE	1	BOROMO	3	C013	173	1
16	1	2	139	2	0.038038	0	200	10	36542	BALE	1	BAGASSI	1	A001	139	2
17	1	2	101	2	0.027639	0	200	10	36542	BALE	1	BAGASSI	1	A002	101	2
18	1	2	184	2	0.050353	0	200	10	36542	BALE	1	BAGASSI	1	A003	184	2
19	1	2	184	2	0.050353	0	200	10	36542	BALE	1	BAGASSI	1	A004	184	2
20	1	2	104	2	0.028460	0	200	10	36542	BALE	1	BAGASSI	1	A005	104	2
21	1	2	259	2	0.070877	0	200	10	36542	BALE	1	BAGASSI	1	B006	259	2
22	1	2	219	2	0.059931	1	200	10	36542	BALE	1	BAGASSI	1	B007	219	2
23	1	2	192	2	0.052542	0	200	10	36542	BALE	1	BAGASSI	1	B008	192	2
24	1	2	224	2	0.061299	0	200	10	36542	BALE	1	BAGASSI	1	B009	224	2
25	1	2	197	2	0.053911	0	200	10	36542	BALE	1	BAGASSI	1	B010	197	2

In Figure 3.9 above, the number of times in which an EA is selected is indicated in the column labeled "# of times selected". Use the filter to locate the selected units and copy them to a new file. Figure 3.10 below gives an example of a portion of a prepared sample file. This is an example; it does not reflect any actual clusters selected for a DHS. The first column gives the cluster number which is assigned by the statistician. The clusters are sorted in the original order as in the sampling frame. The last six columns are the sampling parameters calculated by the program including:

- EA selection probability "Selection Proba",
- number of EAs by stratum "Stratum size",
- number of EAs selected by stratum "Stratum sam-size",
- total measure of size by stratum (total number of households) "Measure size-strat",
- stratum number and
- number of times the unit has been selected.

These are important sampling parameters which must be present in a sample file.

Figure 3.10 Part of an example sample file from a stratified PPS sampling

Cluster number	Province	Code	Commune	Code	EA	HH	type	Selection Proba	Stratum size	Stratum sam-size	Measure size-strat	Stratum number	# of times Select
1	BALE	1	BAGASSI	1	B007	219	2	0.059931	200	10	36542	2	1
2	BALE	1	BAGASSI	1	E026	136	2	0.037217	200	10	36542	2	1
3	BALE	1	BANA	2	B007	301	2	0.082371	200	10	36542	2	1
4	BALE	1	BOROMO	3	B006a	155	1	0.160069	15	3	2905	1	1
5	BALE	1	BOROMO	3	B009	240	1	0.247849	15	3	2905	1	1
6	BALE	1	BOROMO	3	C013	173	1	0.178657	15	3	2905	1	1
7	BALE	1	FARA	4	A001	143	2	0.039133	200	10	36542	2	1
8	BALE	1	FARA	4	E023	193	2	0.052816	200	10	36542	2	1
9	BALE	1	OURY	5	B009	146	2	0.039954	200	10	36542	2	1
10	BALE	1	PA	6	A001	213	2	0.058289	200	10	36542	2	1
11	BALE	1	PA	6	D017	150	2	0.041049	200	10	36542	2	1
12	BALE	1	POURA	8	C011	186	2	0.050900	200	10	36542	2	1
13	BALE	1	YAHO	10	A004	230	2	0.062941	200	10	36542	2	1
14	BANWA	2	BALAVE	11	A002	109	2	0.044718	233	17	41437	4	1
15	BANWA	2	BALAVE	11	D017	209	2	0.085745	233	17	41437	4	1
16	BANWA	2	KOUKA	12	B010	205	2	0.084104	233	17	41437	4	1
17	BANWA	2	KOUKA	12	F027	156	2	0.064001	233	17	41437	4	1
18	BANWA	2	KOUKA	12	I043	117	2	0.048001	233	17	41437	4	1
19	BANWA	2	KOUKA	12	K056	184	2	0.075488	233	17	41437	4	1
20	BANWA	2	SANABA	14	B008	92	2	0.037744	233	17	41437	4	1
21	BANWA	2	SANABA	14	E025	211	2	0.086565	233	17	41437	4	1
22	BANWA	2	SOLENZO	15	A004	93	2	0.038154	233	17	41437	4	1
23	BANWA	2	SOLENZO	15	D019	144	2	0.059078	233	17	41437	4	1
24	BANWA	2	SOLENZO	15	G034	362	2	0.148515	233	17	41437	4	1
25	BANWA	2	SOLENZO	15	J047	240	2	0.098463	233	17	41437	4	1
26	BANWA	2	SOLENZO	15	M062	190	2	0.077950	233	17	41437	4	1
27	BANWA	2	SOLENZO	15	P078	128	1	0.131823	15	3	2913	3	1
28	BANWA	2	SOLENZO	15	Q084	136	1	0.140062	15	3	2913	3	1
29	BANWA	2	SOLENZO	15	R088	274	1	0.282183	15	3	2913	3	1
30	BANWA	2	SOLENZO	15	S090	226	2	0.092719	233	17	41437	4	1
31	BANWA	2	SOLENZO	15	U104	187	2	0.076719	233	17	41437	4	1
32	BANWA	2	TANSILA	16	A005	203	2	0.083283	233	17	41437	4	1
33	BANWA	2	TANSILA	16	D018	233	2	0.095591	233	17	41437	4	1
34	KOSSI	3	BARANI	17	C012	210	2	0.089815	279	20	46763	6	1
35	KOSSI	3	BARANI	17	E026	203	2	0.086821	279	20	46763	6	1
36	KOSSI	3	BARANI	17	G038	158	2	0.067575	279	20	46763	6	1
37	KOSSI	3	BOMBOROKU	18	A004	223	2	0.095375	279	20	46763	6	1
38	KOSSI	3	BOURASSO	19	A002	152	2	0.065009	279	20	46763	6	1
39	KOSSI	3	DJIBASSO	20	A003	234	2	0.100079	279	20	46763	6	1
40	KOSSI	3	DJIBASSO	20	D018	176	2	0.075273	279	20	46763	6	1

3.4 Complex sampling procedures

The sampling procedures used in DHS surveys are usually complex involving multi-stage selection, clustering and stratification, with a combination of *PPS* sampling in the first stage and an equal probability systematic sampling in the second stage. Multi-stage selection is employed due to the lack of a sampling frame at the individual level; clustering is used for implementing efficiency and stratification for the reduction of sampling errors. The DHS sampling procedure has been discussed in some detail in Section 1.8; here we give the basic theoretical properties of the estimator, the variance and variance estimation for a two–stage cluster sampling.

Consider a two-stage stratified cluster sampling, with n_h PSUs selected in stratum h in the first stage with *PPS* sampling, and for each of the selected PSUs, an equal probability systematic sample of m SSUs is selected. Let $y_{hj1}, y_{hj2}, \cdots, y_{hjm}$ be observations from the j^{th} PSU in stratum h. An unbiased estimator of the population total is given by

$$\hat{Y}_{PPS}=\sum_h \sum_{j'} \frac{\hat{Y}_{hj}}{\pi_{Phj}} \quad \text{, with} \quad \hat{Y}_{hj}=\frac{M_{hj}}{m}\sum_{i'} y_{hji}$$

where π_{Phj} is the selection probability of the j^{th} PSU in stratum h; M_{hj} is the number of SSUs in the j^{th} PSU in stratum h. The variance of this estimator is given by

$$V\left(\hat{Y}_{PPS}\right)=\frac{1}{2}\sum_h \sum_{k'} \sum_j \left(\pi_{Phk}\pi_{Phj}-\pi_{Phkj}\right)\left(\frac{Y_{hk}}{\pi_{Phk}}-\frac{Y_{hj}}{\pi_{Phl}}\right)^2 + \sum_h \sum_j \frac{1}{\pi_{Phj}}\frac{1-1/M_{hi}}{m}S_h^2\left(1+(m-1)\rho_{hw}\right)$$

$$\left(V_P\right) \qquad\qquad \left(V_s\right)$$

The first part (V_P) represents the sampling variance of the selection of a PSU, the summation is over all strata for different PSU j and k within the same stratum; the second part (V_S) represents the sampling variance of the selection of an SSU, the summation is over all strata and PSU. Estimators for the first part and second part are obtained from the results in previous sections

$$\hat{V}_P=\frac{1}{2}\sum_h \sum_{k'} \sum_j \frac{\pi_{Phk}\pi_{Phj}-\pi_{Phkj}}{\pi_{Phkj}}\left(\frac{\hat{Y}_{hk}}{\pi_{Phk}}-\frac{\hat{Y}_{hj}}{\pi_{Phl}}\right)^2 \text{¿} \quad , \quad \hat{V}_S=''\sum_h \sum_j \frac{1}{\pi_{Phj}}\frac{1-f_{hj}}{m}s_h^2\text{¿}$$

Since the $\ddot{V}_P$ is not an unbiased estimate of V_P and it usually over estimates V_P, and that V_S is usually smaller compared to V_P, therefore the second part is usually dropped in the variance estimation, this gives an approximate variance estimation given by

$$\hat{V}\left(\hat{Y}_{PPS}\right)=\frac{1}{2}\sum_h \sum_{k'} \sum_j \frac{\pi_{Phk}\pi_{Phj}-\pi_{Phkj}}{\pi_{Phkj}}\left(\frac{\hat{Y}_{hk}}{\pi_{Phk}}-\frac{\hat{Y}_{hj}}{\pi_{Phl}}\right)^2 \text{¿}$$

The above estimator can be simplified as $\hat{V}_R\left(\hat{y}_{PPS}\right)$ in Section 3.3.1

$$\hat{V}_R\left(\hat{Y}_{PPS}\right)=\sum_h \frac{n_h}{n_h-1}\sum_j \left(1-\pi_{Phj}\right)\left(\frac{\hat{Y}_{hj}}{\pi_{Phj}}-\frac{\hat{Y}_h}{n_h}\right)^2 \text{¿}$$

which is reduced to the Woodruff (1971) estimator if $\pi_{Phj}\simeq f_h$ for all h:

$$\hat{V}_W\left(\hat{Y}_{PPS}\right)=\sum_h \frac{n_h\left(1-f_h\right)}{n_h-1}\sum_j \left(\frac{\hat{Y}_{hj}}{\pi_{Phj}}-\frac{\hat{Y}_h}{n_h}\right)^2 \text{¿}$$

where $\hat{Y}_h=\sum_j \frac{\hat{Y}_{hj}}{\pi_{Phj}}$ is the sample estimation of the population total of stratum h.

The above estimator can be expanded to estimate a mean or a ratio by using Woodruff's (1971) linearization approach: let $\hat{R}=\hat{Y}_{PPS}/\hat{X}_{PPS}$, where $\hat{Y}_{PPS}$ represents the total weighted sample value for variable y, and $\hat{X}_{PPS}$ represents the total weighted sample value for variable x or the total number of weighted cases in the group or subgroup under consideration. The approximate variance of $\hat{R}$ can be computed using Woodruff's formula:

$$\hat{V}_W(\hat{R})=\frac{1}{\hat{X}_{PPs}^2}\sum_h \frac{n_h(1-f_h)}{n_h-1}\sum_{i=1}^{n_h}\left(z_{hi}-\frac{z_h}{n_h}\right)^2$$

in which

$$z_{hi}=\left(\hat{Y}_{hj}-\hat{R}\,\hat{X}_{hj}\right)/\pi_{Phj} \quad \text{, and} \quad z_h=\hat{Y}_h-\hat{R}\,\hat{X}_h$$

The above estimator is widely used in commercial statistical software such as SAS, SPSS and Stata. Repeated replication methods such as Bootstrap and Jackknife (Efron, 1982; Efron 1993) can also be used to estimate the variance of $\hat{R}$, as explained in Section 4.2 for estimating sampling errors for complex demographic rates. It should be noted that the DHS survey sampling error calculation procedure has traditionally used the Taylor linearization method (Woodruff, 1971) to calculate the sampling variance for means and ratios because the linearization method is faster computationally than the replication methods.

4 SURVEY ERRORS

The estimates from a sample survey are affected by two types of errors: non-sampling errors and sampling errors. Non-sampling errors are the results of problems occurring during data collection and data processing, such as failure to locate and interview the correct household, misunderstanding of the questions on the part of either the interviewer or the respondent, and data entry errors. Although numerous efforts are made during the implementation of a DHS to minimize this type of error, non-sampling errors are impossible to avoid and difficult to evaluate statistically.

Sampling errors, on the other hand, can be evaluated statistically. The sample of respondents selected in a DHS is only one of many samples that could have been selected from the same population, using the same design and expected size. Each of these samples would yield results that differ somewhat from the results of the actual sample selected. Sampling errors are a measure of the variability between all possible samples. Although the degree of variability is not known exactly, it can be estimated from the survey results. Sampling errors are addressed in some detail in Section 1.6. The following sections of this chapter concentrate on non-sampling errors, including the nature and the sources of errors and the strategies to control them.

As mentioned in Section 1.6, non-sampling errors are usually the main source of errors in a sample survey, and they are difficult to evaluate statistically after the survey is complete. Therefore it is best to minimize this type of error throughout the whole survey implementation process.

4.1 Errors of coverage and non-response

A *coverage error* occurs when a sampling unit is mistakenly excluded from or included in the survey during survey implementation. *Over-coverage* occurs when a non-eligible or a non-sampled sampling unit is deliberately or mistakenly included in the sample; *under-coverage* occurs when a sampled eligible sampling unit is deliberately or mistakenly excluded from the sample. *Non-response*, on the other hand, relates to a failed attempt to interview a sampled sampling unit. This section deals with problems in the definition and estimation of such error rates.

4.1.1 Coverage errors

In DHS surveys, errors of over-coverage (inclusion of units that do not belong in the sample), do not occur as often as under-coverage errors (errors due to exclusion of units that belong in the sample). A typical source of over-coverage occurs when vacant households or non-residential households are sampled for interview. This may occur if a household's occupancy status has changed between the time of the household listing and the household interview. Therefore, it is recommended that the time gap between the household listing and the main data collection should be reasonably small.

For under-coverage, several sources of error may be identified. The first source of under-coverage error arises in the listing stage when the listing staff covers less than the designated area. A second source of under-coverage error occurs when an age limit is used to determine eligibility for individual interview, field staff may misreport an individual's age to push them out of the eligible age range. A third source comes when surveys collect information only from *de facto* individuals (i.e., those who slept in the household the night before the survey). There may be deliberate omissions of eligible individuals by consciously misreporting their "residency" status as non *de facto*, which thereby disqualifies an individual from being eligible for interview. A fourth source comes when a series of questions in the questionnaire are only asked of a certain group. For example, questions related to pregnancy, delivery and child health are only asked for children born since a particular date—there may be omissions of children due to mis-recording of dates of birth as before the cutoff date—or

questions regarding knowledge, attitudes and practices related to HIV are only asked if the respondent is recorded as knowing HIV or AIDS—there may be omissions of respondents due to mis-recording of their knowledge of HIV/AIDS. All four types of coverage errors may involve deliberate bias by fieldworkers seeking to reduce their workload.

Intentional errors can be controlled by intensive training and close supervision. Errors due to an outdated area frame can be reduced by scheduling the household listing operation before the main survey. Errors due to age distortion can be reduced by close supervision and routine quality control. Errors due to residency status can be reduced by changing the data collection strategy to interview all individuals within the age range regardless of their *de facto* status. For example, in DHS surveys, the interviewers are now instructed to interview all women age 15-49 regardless of whether they slept in the household the night before the survey. By requiring the interviewing of all women, the incentive for misreporting residency status has been eliminated. However, the de facto character of the surveys is maintained at the data analysis stage. Using different fieldworkers to conduct the household schedule and individual interviews will also help in eliminating age distortion, misreporting of residency status and mis-recording of dates and other key information. Active monitoring of fieldwork through fieldwork supervision visits and the early use of field check tabulations on collected data can also limit the scope and scale of under-coverage.

Coverage errors can be investigated after the survey fieldwork by a variety of methods. The sample can be extrapolated to the total population, and data from the last census can be extrapolated to the survey date for comparison. This check should be done separately for households and individuals. Age distortions can be investigated by studying the discontinuity in trends across the eligibility boundaries, for example, by looking at the ratio of women age 14 with those age 15, and those age 49 compared with those age 50. While it is tempting to introduce comparisons with males as a control, it should be noted that in most societies more males are educated than females, so more precise knowledge of their own age may reduce heaping at ages 15 and 50 among males compared with females.

4.1.2 Deliberate restrictions of coverage

In many surveys, whether in developed or developing countries, certain parts of the national territory are deliberately excluded from the survey for reasons of difficulty of access. Two distinct cases arise:

- Exclusion of clearly identified areas from the sampling frame—in this case, it is usual to state the coverage limitation in the survey report, which then becomes a report on the remainder of the country. Such exclusions are not regarded as coverage or response errors but simply as part of the definition of the survey domain.

- Ad hoc exclusions decided during or just prior to fieldwork—in many surveys it is not uncommon for the survey organization to abandon the attempt to conduct fieldwork in certain sampled clusters, whether due to floods, civil disturbance, or other practical constraints. Here the exclusions usually occur after sample selection. If such excluded areas form a meaningful domain, it may be acceptable to deal with the problem by redefining the survey domain. More commonly, however, the excluded areas will not form a meaningful domain and will have to be accepted as constituting errors. This type of exclusion should be classified as non-response rather than coverage error.

4.1.3 Non-response

The response rate provides information on the survey coverage problems and is an important survey parameter. At first sight, the concept of non-response seems simple and clear: it occurs when a

sampled unit, household or individual, refuses to be interviewed; the non-response rate is the proportion of the number of non-interviewed units over the number of units selected. Taking into account the distinction between coverage error and non-response indicated earlier, this can be modified by saying that the information desired is the percentage of attempted interviews that failed.

In practice, there are two features found in some sample designs which complicate this simple issue. First, in many surveys the final units for interview are identified through a progressive sifting process. For example, in a typical DHS survey, survey personnel list and select dwellings, interview the household currently in the dwelling, then interview any women age 15-49 in that household. If failure occurs at one of the earlier steps, the information which would enable us to classify the effects at the final level (i.e., the individual level) is lacking. For example, if the interviewer cannot find the selected dwelling, it is not known whether it contains a woman eligible for interview; if the household does not contain any eligible women, then the failure has no effect on the interview response rate.

To deal with this problem, take the women's survey as an example, and assume that there are only two steps in the sifting process, namely households and women. The tradition of DHS surveys is to compute the response rates for the household survey and the women's interview separately because of the way that sample weights are calculated. There are six quantities of potential interest in computing response rates:

A. Households selected

B. Households found or eligible (excluding vacant, destroyed, etc.)

C. Households interviewed

D. Women selected

E. Women found or eligible (all *de facto* women 15-49 found)

F. Women interviewed

Since the survey primarily concerns women, the relevant response rate is F/D (i.e., women interviewed divided by women selected). However, the quantity D is unknown because of the non-responding households. It is of interest to know the total number of eligible women in all selected households but, only the number the number of women found in the households interviewed (E) is known. Therefore D must be estimated by taking the household non-response into account. Assuming that the number of eligible women per household is the same among non-responding households as it is among interviewed households, the number of women selected can be estimated as:

$$D = E \div \frac{C}{B}$$

where C/B is the effective household response rate. The reason to use the effective household response rate is that the non-eligible (vacant, destroyed or other) households A-B is considered as over-coverage, assuming that same over-coverage exists in the household listing. These assumptions may not be very convincing, but the effect of any departure from them on the estimate of D is likely to be very small. On this basis the overall response rate for the women's survey, R=F/D, becomes:

$$R = \frac{F}{D} = \frac{F}{E} \times \frac{C}{B}$$

This response rate is the product of the response rates observed at each of the two stages, households and women. This basic principle provides a solution for the problem of not knowing the

total number of women sampled. Where two or more steps of sifting are involved, the overall response rate can be estimated by multiplying together the response rates observed at each step. In doing so, the assumption is made that the response/non-response outcomes at the different steps occur independently.

DHS surveys do not allow the replacement of non-responding households because of the potential bias which may result from the replaced households being easier to contact. However, when a sampled household in a selected dwelling moves away between the listing and the interview, the MEASURE DHS program recommends interviewing the new household (if any) that has moved in by the time of the main survey. This is not considered a replacement; in fact it reflects the fact that the sampling unit is defined as the dwelling structure rather than its occupants. The design calls for the listing and selection of dwellings, and then for the interview of the household found in the dwelling at the time of the survey. Since in many areas there is no address system, the initial listing operation has to identify the dwellings in terms of the names of the occupying households, but these merely serve as addresses. The fact that, in some cases, a new household moves in between the time of listing and interview does not mean that replacement of a sampling unit has occurred. Thus, such cases do not require any special treatment. Moreover, just as a new household moving in does not constitute a replacement, so the case of a household moving out after the listing without another moving in, creating a vacant household, does not constitute non-response. The eligible household sample is defined as the set of households existing at the time of interviewing in the dwellings selected from the dwelling list.

4.1.4 Response rates

As seen in the previous section, the women's overall response rate is the product of the observed household and women's response rates, therefore, it is meaningful to calculate these two response rates separately. As we mentioned in Section 1.13, non-response brings bias. Therefore, the different response rates reflect the data quality. A separate response rate is useful in sample size design and field work improvement. In order to categorize in detail the non-responding households and individuals, the MEASURE DHS program standardized the response codes to be entered on the questionnaires and field records, and expressed the formulae for response rates in terms of these codes. In DHS surveys, the following response categories are used at the household level:

1H Completed

2H No household member at home or no competent respondent at home

3H Entire household absent for extended period

4H Postponed

5H Refused

6H Dwelling vacant or address not a dwelling

7H Dwelling destroyed

8H Dwelling not found

9H Other

Note that household above refers to the household found in the dwelling at the time of the interview, not necessarily the household named at the time of the listing operation. The DHS survey final reports provide the household response rate calculated by:

$$R_H = \frac{1H}{1H + 2H + 4H + 5H + 8H}$$

The reason to include 8H in the denominator is that a household that is not found at the time of the fieldwork may not be a vacant household. It may be that the household was not found because of some error that occurred during the survey implementation. Note also that this response rate is different from the weighted response rate calculated in Section 1.13. In Section 1.13 the aim is to calculate the sampling weight, while here the response rate is used as a data quality indicator. It is also worth noting that the above calculated response rate is a *net response rate*. For the purpose of sample size determination, one should use the *gross response rate* which is the number of households interviewed over the number selected:

$$R_{HG} = \frac{1H}{1H + 2H + 3H + 4H + 5H + 6H + 7H + 8H + 9H}$$

If the net response rate is used to calculate sample size, the survey may not obtain the designed number of interviews because some of sampled households will always end up being non-eligible, especially when there is a long time lag between household listing and the main field work.

At the individual level the following response categories are used:

1I Completed

2I Not at home

3I Postponed

4I Refused

5I Partly completed

6I Incapacitated

7I Other

The individual response rate is thus:

$$R_I = \frac{1I}{1I + 2I + 3I + 4I + 5I + 6I + 7I}$$

The category “no eligible woman in the household” is not included in the list since it is irrelevant to the response rate, appearing neither in the numerator nor the denominator. The same is true for “non *de facto* women.” Although an individual questionnaire is administered to non-*de facto* women who live in the household to reduce under-coverage errors as mentioned in Section 4.1.1, these interviews are not counted in the numerator or the denominator of the response rate because non-*de facto* women are not eligible according to the definition of eligibility.

Whenever the *other* code is used, the interviewers should specify the reason for non-response. At the household level, the analyst should review a printout of the other codes and recode as many as possible into the existing categories. Similarly, all other codes for the individual interview should be examined and recoded. Any questionnaire in which the household or the woman was deemed ineligible should be clearly marked as ineligible and removed from the data file. An ineligible household may be one in a dwelling unit that does not lie within the sample area or a neighboring household that was interviewed incorrectly as a replacement household. An ineligible woman may be one who was

reported as 16 years old in the household questionnaire, but later turned out to be 14 (in which case her age in the household questionnaire should be corrected appropriately).

The overall response rate is obtained by multiplying the household and the individual level response rates:

$$R = R_h \times R_I$$

However, if there has been a deliberate exclusion of certain areas such as clusters which were not interviewed (see Section 1.13 on cluster level non-response), the overall response rate must also take the cluster response rate into account. In summary, the final overall estimated response rate is obtained from the formula:

$$R = R_h \times R_I \times R_C$$

where $R_c = n^{'} / n$ is the ratio of the number of clusters interviewed over the number selected.

Such response rates should be computed and published separately for the main geographic domains of the sample as well as the whole survey domain. If the sample is self-weighting within domain but has different weights across domains, the response rates should be computed and published for each differently weighted domain.

4.2 Sampling errors

We introduced the concept of sampling errors in Section 1.6 for sample size determination. In this section, we focus on the calculation of the sampling errors. Sampling errors are usually reported for selected indicators in Appendix B of the DHS final report.

A sampling error is usually measured in terms of the standard error for a particular statistic (mean, percentage, etc.), which is the square root of the variance. The standard error can be used to calculate confidence intervals within which the true value for the population can reasonably be assumed to fall. For example, for any given statistic calculated from a sample survey, the value of that statistic will fall within a range of plus or minus two times the standard error (DHS reports +/-2*SE instead of +/-1.96*SE as 95% confidence interval as explained in section 1.6.1) of that statistic in 95 percent of all possible samples of identical size and design.

If the sample of respondents were selected as a simple random sample, it would have been possible to use straightforward formulae to calculate sampling errors. However, DHS survey samples are the result of a multi-stage stratified design, so it is necessary to use more complex formulae. There is a variety of computer software which can be used to calculate sampling errors, such as the Integrated System for Survey Analysis (ISSA) sampling errors module and the ICF developed SAS macro as well as software such as Wesvar, Cenvar, and Sudaan. These software use the Taylor Linearization Method (Woodruff, 1971) of variance estimation for survey estimates that are means or proportions. This same method is widely used in commercialized statistical software such as SAS, SPSS and STATA. The Jackknife Repeated Replication Method (Efron, 1982, 1993) is used for variance estimation of more complex statistics such as fertility and mortality rates.

The Taylor Linearization Method treats any percentage or average as a ratio estimate, $r = y/x$, where y represents the total weighted sample value for variable y, and x represents the total weighted sample value for variable x or the total number of weighted cases in the group or subgroup under consideration. The variance of r is computed using the formula given below, with the standard error being the square root of the variance:

$$SE^2(r) = var(r) = \frac{1}{x^2} \sum_{h=1}^{H} \frac{n_h(1-f_h)}{n_h - 1} \sum_j \left(z_{hj} - \frac{z_h}{n_h} \right)^2$$

in which

$$z_{hi} = y_{hi} - rx_{hi} \quad \text{, and} \quad z_h = y_h - rx_h$$

where h represents the sampling stratum which varies from 1 to H,
n_h is the total number of clusters selected in the h^{th} stratum,
y_{hj} is the sum of weighted values of variable y in the j^{th} cluster in the h^{th} stratum,
x_{hj} is the sum of weighted values of variable x in the j^{th} cluster in the h^{th} stratum,
f_h is the sampling fraction in stratum h, it can be ignored when it is small
x is the sum of weighted values of variable x over the total sample

The Jackknife Repeated Replication Method derives estimates of complex rates from each of several replications of the parent sample, and calculates standard errors for these estimates using simple formulae. Each replication considers all but one cluster in the calculation of the estimates. Pseudo-independent replications are thus created. The variance of a rate r is calculated as follows:

$$SE^2(r) = Var(r) = \frac{1}{k(k-1)} \sum_{i=1}^{k} (r_i - r)^2$$

in which

$$r_i = kr - (k-1) r_{(i)}$$

where r is the estimate computed from the full sample of k clusters,
$r_{(i)}$ is the estimate computed from the reduced sample of *k-1* clusters (with i^{th} cluster excluded), and
k is the total number of clusters.

In addition to the standard error, the procedure computes the design effect (*DEFT*) for estimates which are means, proportions or ratios. For complex demographic rates, the procedure computes an approximation of *DEFT*. *DEFT* is defined as the ratio between the standard error using the given sample design and the standard error that would result if a simple random sample had been used. A *DEFT* value of 1.0 indicates that the sample design is as efficient as a simple random sample, while a value greater than 1.0 indicates the increase in the sampling error due to the use of a more complex and less statistically efficient design. The procedure also computes the relative error and confidence limits for the estimates.

Sampling errors are usually reported for the total sample, for the urban and rural areas, and for each of the survey domains.

5 SAMPLE DOCUMENTATION

5.1 Introduction

Sample documentation is an important part of a DHS survey. The documentation should include all useful information for data analysis, for data quality assessment, for sample design of subsequent surveys, and for data users. Basic sample documentation should be included in DHS survey final reports. Good sample documentation should include the following aspects from different stages of the survey implementation:

1) Target population
2) Expected sample size
3) Main indicators
4) Report domains
5) Sampling frame
6) Primary and the secondary sampling units
7) Stratification
8) Sample allocation
9) Sampling procedure
10) Selection probability
11) Household listing results
12) Sampling weights
13) Results of survey implementation
14) Sampling errors

Points 1 to 10 and point 12 are usually addressed in a Sample Design Document from the very beginning of the survey. For point 11, the number of households listed, the number of households selected, and segmentation information for each of the selected clusters should be provided. A full description sample design should be included in Appendix A of the DHS final reports. For point 13, the number of eligible sampling units selected, the number interviewed and the household and individual response rates should be presented. Sampling errors (point 14) are presented in Appendix B of DHS final reports for selected indicators.

5.2 Sample design document

A sample design document is an important document which records the purpose of the survey, the target population, the source of the sampling frame, the statistical methodology, the sample size and the sample allocation, and other related topics. This section gives an example of a sample design document to show the details which should be included in a sample design document.

5.2.1 Introduction

The Country Demographic and Health Survey 2012 (XDHS 2012) will be the fourth DHS following those implemented in 1995, 2000 and 2005. A nationally representative sample of 18,450 households will be selected. All women 15-49 who are usual residents of a selected household or who slept in a selected household the night before the survey are eligible for the survey. The survey will result in about 17,900 interviews of women 15-49. As with the prior surveys, the main objectives of the XDHS 2012 survey are to provide up-to-date information on fertility and childhood mortality levels; fertility preferences; awareness, approval and use of family planning methods; maternal and child health; knowledge and attitudes toward HIV/AIDS and other sexually transmitted infections (STI).

Apart from the women's survey, a men's survey will also be conducted at the same time in a sub-sample consisting of one household in every three selected for the women's survey. All men 15-59 who are usual residents of a selected household or who slept in a selected household the night before the survey are eligible for the men's survey. The survey will collect information on their basic demographic and social status; on their knowledge and use of family planning methods; and on their knowledge and attitudes toward HIV/AIDS and other sexually transmitted infections. The survey will result in about 5,000 interviews of men 15-49. In this sub-sample, all women 15-49, all children under 5 years of age will be weighed, measured and tested for anemia in order to study their nutritional status.

The survey is designed to produce representative estimates for most of the indicators for the country as a whole, for the urban and the rural areas separately, for the capital city of the country, and for each of the ten geographical regions.

5.2.2 Sampling frame

The sampling frame used for XDHS 2010 is the Country Population and Housing Census conducted in 2006 (XPHC 2006), provided by the Central Statistical Office (CSO). CSO has made available an electronic file consisting of 81,654 Enumeration Areas (EAs) created for the 2006 census in 9 of its 10 regions. An EA is a geographic area consisting of a convenient number of dwelling units which served as a counting unit for the census. The frame file contains information about the location, the type of residence and the number of residential households for each of the 81,654 EAs. Sketch maps are also available for each EA which delineate the geographic boundaries of the EA. It should be pointed out that this file does not include Region 10 because the census conducted in Region 10 used a different methodology due to difficulty of access. Therefore, the sampling frame for Region 10 is in a different file and uses a different format. It is also worth noting that the sampling frame excluded some special EAs which have disputed boundaries; this kind of EA represents only 0.1% of the total population.

The census cartographic work for Region 10 was conducted using two different methods. In two of its six districts, namely, Districts 2 and 4, traditional cartographical work similar to the other regions of the country was carried out, while in the other four districts, the cartographic work was carried out by using satellite photos without physical visits of the area. The census data could not be used to update the cartographic work in Region 10 because of coding problems. So in Region 10, a sampling frame with a similar format as in the other regions is available only for the three zones where a traditional cartographic work had been carried out. However, the number of households in the sampling frame for these three zones is based on the number of households estimated during the cartographic work preceding the census and not the actual number of households counted in the census. Due to security concerns, as in the XDHS 2000 and XDHS 2005, it has been decided that the XDHS 2012 will be conducted only in these two districts. These two districts together have 1,246 EAs, and they represent 53% of the regional total population. Taking into account the special EAs which are excluded from the census frame, the sampling frame used for the XDHS 2012 covered 98.4% of the country's total population.

Country is divided into 10 geographical regions; each region is sub-divided into districts, and each districts into wards. Table 5.1 shows the distribution of the EAs and the mean number of households per EA by region and by type of residence. The sampling frame includes 82,900 EAs, among them 17,346 are in urban areas and 65,554 are in rural areas. The average size of an EA in terms of number of households is 170 in an urban EA and 182 in a rural EA, for an overall average size of 180 households per EA. Table 5.2 shows the distributions of households by region and by type of residence. The distribution is a very skewed distribution since 83.4% of the country's households are concentrated in 3 regions, namely, Region 3, Region 4 and Region 6; while the five small regions

Region 2, Region 5, Region 7, Region 8 and Region 9 together represent only 3.8% of the country's total households.

Table 5.1 Distribution of EAs and average size of EA by region and by type of residence

Region	Number of EA			Average EA size		
	Urban	Rural	Total	Urban	Rural	Total
Region 1	1,541	4,139	5,680	153	177	171
Region 2	260	828	1,088	177	233	219
Region 3	3,391	18,016	21,407	183	182	182
Region 4	5,030	25,800	30,830	172	179	178
Region 5	188	786	974	140	152	150
Region 6	2,124	14,490	16,614	166	184	182
Region 7	133	347	480	145	129	134
Region 8	172	98	270	163	180	169
Capital City	3,865		3,865	167		167
Region 9	318	128	446	163	169	165
Region 10*	324	922	1,246	154	267	237
Country	**17,346**	**65,554**	**82,900**	**170**	**182**	**180**

Source: XPHC 2006; Region 10 has only two districts included.

Table 5.2 Distribution of households by region and by type of residence

Region	Number of households			% Urban	% of Country
	Urban	Rural	Total		
Region 1	235,530	734,357	969,887	0.243	0.065
Region 2	45,910	192,554	238,464	0.193	0.016
Region 3	619,796	3,284,512	3,904,308	0.159	0.262
Region 4	864,303	4,630,702	5,495,005	0.157	0.369
Region 5	26,314	119,446	145,760	0.181	0.010
Region 6	353,554	2,667,787	3,021,341	0.117	0.203
Region 7	19,275	44,879	64,154	0.300	0.004
Region 8	27,975	17,651	45,626	0.613	0.003
Capital City	646,216	0	646,216	1.000	0.043
Region 9	51,991	21,643	73,634	0.706	0.005
Region 10*	49,844	245,922	295,766	0.169	0.020
Country	**2,940,708**	**11,959,453**	**14,900,161**	**0.197**	**1.000**

Source: XPHC 2006; Region 10 has only two districts included.

5.2.3 Structure of the sample and the sampling procedure

The sample for the XDHS 2012 will be a stratified sample selected in two stages from the 2006 census frame. Stratification was achieved by separating each region into urban and rural areas. In total, 19 sampling strata have been created since the region of Capital has only urban areas. Samples will be selected independently in each sampling stratum, by two-stage selection. Implicit stratification and proportional allocation is achieved at each of the lower administrative levels by sorting the

sampling frame according to administrative units in different levels and by using a probability proportional to size selection at the first stage of sampling.

In the first stage, 615 EAs have been selected with probability proportional to EA size and with independent selection in each sampling stratum with the sample allocation given in table 5.3 below. Taking into account the time passed since the last population census, a household listing operation will be carried out in all of the selected EAs before the main survey. The household listing operation consists of visiting each of the 615 selected EAs; drawing a location map and a detailed sketch map; and recording on the household listing forms all residential households found in the EA with the address and the name of the head of the household. The resulting list of households will serve as the sampling frame for the selection of households in the second stage. Some of the selected EAs may be found to be large in size in the household listing operation. In order to minimize the task of household listing, the selected EAs containing an estimated number of households greater than 300 will be segmented. Only one segment will be selected for the survey with probability proportional to the segment size. The methodology and the detailed household listing procedure are addressed in the Household Listing Manual (see Chapter 2).

At the second stage, a fixed number of 30 households will be selected from each EA. Table 5.3 shows the sample distribution of clusters and households by region and by type of residence. Among the 615 EAs selected, 185 are in urban areas and 430 are in rural areas. The total number of households to be selected is 18,450; among them, 5,550 will be in urban areas and 12,900 will be in rural areas.

In the sampling frame, the household distribution by region varies from 0.3 percent for Region 8, to 36.9 percent for Region 4 (see Table 5.2 in Section 5.2.2). To allocate the approximately 17,900 women interviews to different regions, a proportional allocation will provide the best precision for national level indictors, but not for regional level indicators. The small regions such as Region 7, Region 8 and Region 9 would receive a sample size which is too small to achieve the degree of precision desired for regional level estimates. In order for the precision of estimates to be acceptable across regions, experience shows that a minimum of 800 women’s interviews are needed so that reliable estimations for most of the DHS indicators can be obtained. The final sample allocation reflects a power allocation which is between the proportional allocation and the equal size allocation. So that the survey precision in the urban areas is comparable with the rural areas, urban areas are slightly over-sampled.

The allocations of clusters and households by region and by type of residence are functions of the estimated average number of women age 15-49 per household and the household and individual response rates. Estimates for these parameters are obtained from the XDHS 2005 survey. According to the results of XDHS 2005, the average number of women age 15-49 per household is 1.20 in urban areas and 1.00 in rural areas. The number of men age 15-49 per household is 1.05 in urban areas and 0.95 in rural areas. The household response rates are 92 percent in urban areas and 94 percent in rural areas; the women’s response rates are 94 percent and 96 percent in the urban and rural areas, respectively; the men’s response rates are 85 percent and 90 percent in the urban and rural areas, respectively.

Table 5.3 Sample allocation of clusters and households by region and by type of residence

Region	Allocation of clusters			Allocation of households		
	Urban	Rural	Region	Urban	Rural	Region
Region 1	13	47	60	390	1,410	1,800
Region 2	10	38	48	300	1,140	1,440
Region 3	10	62	72	300	1,860	2,160
Region 4	13	62	75	390	1,860	2,250
Region 5	6	42	48	180	1,260	1,440
Region 6	7	65	72	210	1,950	2,160
Region 7	9	37	46	270	1,110	1,380
Region 8	25	17	42	750	510	1,260
Capital City	54	na	54	1,620	na	1,620
Region 9	27	15	42	810	450	1,260
Region 10	11	45	56	330	1,350	1,680
Country	**185**	**430**	**615**	**5,550**	**12,900**	**18,450**

Table 5.4 Expected number of interviews by region and by type of residence

Statistical Region	Women interviewed			Men interviewed		
	Urban	Rural	Region	Urban	Rural	Region
Region 1	434	1,280	1,714	98	358	456
Region 2	333	1,035	1,368	76	290	366
Region 3	333	1,689	2,022	76	472	548
Region 4	434	1,689	2,123	98	472	570
Region 5	200	1,144	1,344	45	320	365
Region 6	233	1,771	2,004	53	495	548
Region 7	299	1,008	1,307	69	282	351
Region 8	834	463	1,297	189	130	319
Capital City	1,800	na	1,800	408	na	408
Region 9	901	409	1,310	205	114	319
Region 10	367	1,226	1,593	83	342	426
Country	**6,168**	**11,714**	**17,882**	**1,400**	**3,275**	**4,676**

Men's survey will be carried out in one household in every three selected for women's survey.

5.2.4 Selection probability and sampling weight

Due to the non-proportional allocation of the sample to the different regions and to their urban and rural areas, sampling weights will be required for any analysis using XDHS 2012 data to ensure the survey results are representative at national and regional levels. Since the XDHS 2012 sample is a two-stage stratified cluster sample, sampling weights will be calculated based on the separate sampling probabilities for each sampling stage and for each cluster. We use the following notations:

P_{1hi}: first-stage sampling probability of the i^{th} cluster in stratum h

P_{2hi}: second-stage sampling probability within the i^{th} cluster (household selection)

Let n_h be the number of clusters selected in stratum h, M_{hi} the number of households according to the sampling frame in the i^{th} cluster, and $\sum M_{hi}$ the total number of households in the stratum. The probability of selecting the i^{th} cluster in the XDHS 2012 sample is calculated as follows:

$$P_{1hi} = \frac{n_h M_{hi}}{\sum M_{hi}}$$

A different formula must be used to calculate the probability of selecting a cluster that has been segmented. Let b_{hi} be the proportion of households in the selected segment compared to the total number of households in the EA i in stratum h if the EA is segmented, otherwise $b_{hi}=1$. Then the probability of selecting cluster i in the sample is:

$$P_{1hi} = \frac{n_h M_{hi}}{\sum M_{hi}} \times b_{hi}$$

Let L_{hi} be the number of households listed in the household listing operation in cluster i in stratum h, let t_{hi} be the number of households selected in the cluster. The second stage selection probability for each household in the cluster is calculated as follows:

$$P_{2hi} = \frac{t_{hi}}{L_{hi}}$$

The overall selection probability of each household in cluster i of stratum h is therefore the production of the two selection probabilities:

$$P_{hi} = P_{1hi} \times P_{2hi}$$

The design weight for each household in cluster i of stratum h is the inverse of its overall selection probability:

$$W_{hi} = 1/P_{hi}$$

A spreadsheet containing all sampling parameters and selection probabilities is prepared to facilitate the calculation of sampling weights. Sampling weights will be adjusted for household non-response as well as for individual non-response, for the women's and men's surveys respectively. The differences between the household weights and the individual weights are introduced by individual non-response. The final weights are normalized so that the total number of unweighted cases will equal the total number of weighted cases at the national level, for both household weights and individual weights.

5.3 Sample file

A sample file including all sampling parameters is very important for survey management and for sampling weight calculation. Once the sample points are selected, an Excel file should be prepared which should include the cluster number and cluster ID information, and all sampling parameters such as the domain, stratum and EA selection probability. The cluster number is a unique serial number

from 1 to the total number of clusters selected. It is important for communication and for field work supervision. The cluster number is the official cluster ID once assigned. It is also useful to include in the sample file the EA size, the total size of the stratum, the number of EAs in the stratum and the number of EAs selected in the stratum. These pieces of information allow for reconstruction of the selection probability, if needed, for example, for checking purposes and for replacement clusters.

If a selected cluster is not accessible due to security problems and a replacement cluster is selected, then from the sampling parameters it is easy to calculate the selection probability for the replacement cluster. Table 5.5 below shows a part of an example sample file. The columns with the lighter colored headings represent the sampling information provided by the sampling statistician. The columns with the darker colored headings represent the EA identification information from the sampling frame. This file should be updated after the household listing operation by adding the number of households listed, the segmentation information, and the number of households selected. These 3 pieces of information are necessary for developing the design weight for each cluster.

Table 5.5 An example sample file

Cluster number	Province	Code	Commune	Code	EA	HH	Urban=1 /Rural=2	Selection Proba	Stratum size	Stratum sam-size	Measure size-strat	Stratum number	# of times Select
1	Province 1	1	Commune 1	1	B007	219	2	0.059931	200	10	36542	2	1
2	Province 1	1	Commune 1	1	E026	136	2	0.037217	200	10	36542	2	1
3	Province 1	1	Commune 2	2	B007	301	2	0.082371	200	10	36542	2	1
4	Province 1	1	Commune 3	3	B006a	155	1	0.160069	15	3	2905	1	1
5	Province 1	1	Commune 3	3	B009	240	1	0.247849	15	3	2905	1	1
6	Province 1	1	Commune 3	3	C013	173	1	0.178657	15	3	2905	1	1
7	Province 1	1	Commune 4	4	A001	143	2	0.039133	200	10	36542	2	1
8	Province 1	1	Commune 4	4	E023	193	2	0.052816	200	10	36542	2	1
9	Province 1	1	Commune 5	5	B009	146	2	0.039954	200	10	36542	2	1
10	Province 1	1	Commune 6	6	A001	213	2	0.058289	200	10	36542	2	1
11	Province 1	1	Commune 6	6	D017	150	2	0.041049	200	10	36542	2	1
12	Province 1	1	Commune 8	8	C011	186	2	0.050900	200	10	36542	2	1
13	Province 1	1	Commune 10	10	A004	230	2	0.062941	200	10	36542	2	1
14	Province 2	2	Commune 11	11	A002	109	2	0.044718	233	17	41437	4	1
15	Province 2	2	Commune 11	11	D017	209	2	0.085745	233	17	41437	4	1
16	Province 2	2	Commune 12	12	B010	205	2	0.084104	233	17	41437	4	1
17	Province 2	2	Commune 12	12	F027	156	2	0.064001	233	17	41437	4	1
18	Province 2	2	Commune 12	12	I043	117	2	0.048001	233	17	41437	4	1
19	Province 2	2	Commune 12	12	K056	184	2	0.075488	233	17	41437	4	1
20	Province 2	2	Commune 14	14	B008	92	2	0.037744	233	17	41437	4	1
21	Province 2	2	Commune 14	14	E025	211	2	0.086565	233	17	41437	4	1
22	Province 2	2	Commune 15	15	A004	93	2	0.038154	233	17	41437	4	1
23	Province 2	2	Commune 15	15	D019	144	2	0.059078	233	17	41437	4	1
24	Province 2	2	Commune 15	15	G034	362	2	0.148515	233	17	41437	4	1
25	Province 2	2	Commune 15	15	J047	240	2	0.098463	233	17	41437	4	1
26	Province 2	2	Commune 15	15	M062	190	2	0.077950	233	17	41437	4	1
27	Province 2	2	Commune 15	15	P078	128	1	0.131823	15	3	2913	3	1
28	Province 2	2	Commune 15	15	Q084	136	1	0.140062	15	3	2913	3	1
29	Province 2	2	Commune 15	15	R088	274	1	0.282183	15	3	2913	3	1
30	Province 2	2	Commune 15	15	S090	226	2	0.092719	233	17	41437	4	1
31	Province 2	2	Commune 15	15	U104	187	2	0.076719	233	17	41437	4	1
32	Province 2	2	Commune 16	16	A005	203	2	0.083283	233	17	41437	4	1
33	Province 2	2	Commune 16	16	D018	233	2	0.095591	233	17	41437	4	1
34	Province 3	3	Commune 17	17	C012	210	2	0.089815	279	20	46763	6	1
35	Province 3	3	Commune 17	17	E026	203	2	0.086821	279	20	46763	6	1
36	Province 3	3	Commune 17	17	G038	158	2	0.067575	279	20	46763	6	1
37	Province 3	3	Commune 18	18	A004	223	2	0.095375	279	20	46763	6	1
38	Province 3	3	Commune 19	19	A002	152	2	0.065009	279	20	46763	6	1
39	Province 3	3	Commune 20	20	A003	234	2	0.100079	279	20	46763	6	1
40	Province 3	3	Commune 20	20	D018	176	2	0.075273	279	20	46763	6	1
41	Province 3	3	Commune 20	20	F032	187	2	0.079978	279	20	46763	6	1
42	Province 3	3	Commune 20	20	I046	366	2	0.156534	279	20	46763	6	1
43	Province 3	3	Commune 20	20	K059	265	2	0.113337	279	20	46763	6	1
44	Province 3	3	Commune 21	21	C017	195	2	0.083399	279	20	46763	6	1
45	Province 3	3	Commune 21	21	F033	225	2	0.096230	279	20	46763	6	1
46	Province 3	3	Commune 22	22	D018	149	2	0.063726	279	20	46763	6	1
47	Province 3	3	Commune 22	22	F032	203	2	0.086821	279	20	46763	6	1
48	Province 3	3	Commune 24	24	A002	219	2	0.093664	279	20	46763	6	1
49	Province 3	3	Commune 25	25	B009	234	2	0.100079	279	20	46763	6	1
50	Province 3	3	Commune 25	25	D021	217	2	0.092808	279	20	46763	6	1
51	Province 3	3	Commune 25	25	H038	205	2	0.087676	279	20	46763	6	1
52	Province 3	3	Commune 25	25	K052	237	2	0.101362	279	20	46763	6	1
53	Province 3	3	Commune 25	25	M061	212	1	0.176716	19	3	3599	5	1
54	Province 3	3	Commune 25	25	N068	159	1	0.132537	19	3	3599	5	1
55	Province 3	3	Commune 25	25	O074	187	1	0.155877	19	3	3599	5	1
56	Province 3	3	Commune 26	26	B008	157	2	0.067147	279	20	46763	6	1

5.4 Results of Survey implementation

Once the field work for the survey has been completed, and the data entry is finished, some tables for the results of the survey implementation should be produced to evaluate the survey coverage and the departures from the survey design. These tables typically include a summary table and individual tables for the household, women's and men's surveys, respectively. A summary table is usually presented in Chapter 1 of the DHS final report, including the number of clusters selected and interviewed, the number of households selected and interviewed, the number of women selected and interviewed, and the number of men selected and interviewed. The detailed tables for the household, women's and men's surveys are usually present in Appendix A of the DHS final report along with the sample design document. These tables both reflect the survey coverage and the data quality and provide various response rates and the number of eligible individuals per household, which are useful information for the sample design for subsequent surveys. The following tables are example tables that should be included in the final report.

Table 5.6 Example table for the results of survey implementation

Table 1.1 Results of the household and individual interviews

Number of households, number of interviews, and response rates, according to residence (unweighted), Bangladesh 2007

	Residence					
	Urban		Rural		Total	
Result	Number	Percent	Number	Percent	Number	Percent
Households selected	3,993	100.0	6,826	100.0	10,819	100.0
Households occupied	3,849	96.4	6,612	96.9	10,461	96.7
Households absent for extended period	78	2.0	121	1.8	199	1.8
Dwelling vacant or destroyed	59	1.5	73	1.1	132	1.2
Other	7	0.2	20	0.3	27	0.2
Household interviews						
Households occupied	3,849	96.4	6,612	96.9	10,461	96.7
Households interviewed	3,821	95.7	6,579	96.4	10,400	96.1
Household response rate[1]		99.3		99.5		99.4
Interviews with women age 15-49						
Number of eligible women	4,230	100.0	6,948	100.0	11,178	100.0
Number of eligible women interviewed	4,151	98.1	6,845	98.5	10,996	98.4
Eligible women response rate[2]		98.1		98.5		98.4
Interviews with men age 15-54						
Number of eligible men	1,559	100.0	2,515	100.0	4,074	100.0
Number of eligible men interviewed	1,443	92.6	2,328	92.6	3,771	92.6
Eligible men response rate[2]		92.6		92.6		92.6

[1] Households interviewed/households occupied
[2] Respondents interviewed/eligible respondents

Table 5.7 Example appendix table for the results of the women's survey implementation

Table A.5 Sample implementation: Women

Percent distribution of households and eligible women by results of the household and individual interviews, and household, eligible women and overall response rates, according to urban-rural residence and region, Bangladesh 2007

	Residence		Division						
Result	Urban	Rural	Barisal	Chittagong	Dhaka	Khulna	Rajshahi	Sylhet	Total
Selected households									
Completed (C)	95.7	96.4	96.3	96.5	95.0	96.1	96.8	96.4	96.1
Household present but no competent respondent at home (HP)	0.5	0.4	0.7	0.4	0.6	0.5	0.1	0.4	0.5
Refused (R)	0.2	0.0	0.0	0.1	0.1	0.1	0.0	0.0	0.1
Dwelling not found (DNF)	0.0	0.1	0.0	0.1	0.0	0.0	0.0	0.1	0.0
Household absent (HA)	2.0	1.8	1.8	1.7	2.4	1.4	1.9	1.8	1.8
Dwelling vacant/address not a dwelling (DV)	1.3	0.9	0.9	0.9	1.5	1.5	0.7	0.8	1.1
Dwelling destroy (DD)	0.2	0.2	0.2	0.3	0.0	0.2	0.1	0.2	0.2
Other (O)	0.2	0.3	0.1	0.1	0.4	0.2	0.3	0.3	0.2
Total	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Number of sampled households	3,993	6,826	1,470	1,860	2,340	1,680	2,070	1,399	10,819
Household response rate (HRR)[1]	99.3	99.5	99.2	99.4	99.2	99.4	99.8	99.4	99.4
Eligible women									
Completed (EWC)	98.1	98.5	98.0	98.0	97.7	99.2	99.2	98.2	98.4
Not at home (EWNH)	1.3	1.1	1.6	1.2	2.0	0.6	0.4	1.2	1.2
Postponed (EWP)	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0
Refused (EWR)	0.2	0.0	0.0	0.3	0.1	0.0	0.0	0.0	0.1
Partly completed (EWPC)	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.1	0.0
Incapacitated (EWI)	0.3	0.3	0.3	0.3	0.3	0.2	0.2	0.5	0.3
Total	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Number of women	4,230	6,948	1,467	1,983	2,396	1,725	2,096	1,511	11,178
Eligible women response rate (EWRR)[2]	98.1	98.5	98.0	98.0	97.7	99.2	99.2	98.2	98.4
Overall response rate (OWRR)[3]	97.4	98.0	97.3	97.4	96.9	98.6	99.0	97.6	97.8

[1] Using the number of households falling into specific response categories, the household response rate (HRR) is calculated as:

$$\frac{100 * C}{C + HP + P + R + DNF}$$

[2] Using the number of eligible women falling into specific response categories, the eligible women response rate (EWRR) is calculated as:

$$\frac{100 * EWC}{EWC + EWNH + EWP + EWR + EWPC + EWI + EWO}$$

[3] The overall women response rate (OWRR) is calculated as:

$$OWRR = HRR * EWRR/100$$

Table 5.8 Example appendix table for the results of the men's survey implementation

Table A.6 Sample implementation: Men

Percent distribution of households and eligible men by results of the household and individual interviews, and household, eligible men and overall response rates, according to urban-rural residence and region, Bangladesh 2007

	Residence		Division						
Result	Urban	Rural	Barisal	Chittagong	Dhaka	Khulna	Rajshahi	Sylhet	Total
Selected households									
Completed (C)	95.3	96.6	96.1	96.6	94.0	96.0	97.2	97.6	96.1
Household present but no competent respondent at home (HP)	0.5	0.4	0.5	0.5	0.7	0.7	0.0	0.3	0.5
Refused (R)	0.2	0.0	0.0	0.2	0.1	0.1	0.0	0.0	0.1
Dwelling not found (DNF)	0.1	0.1	0.0	0.0	0.1	0.0	0.1	0.1	0.1
Household absent (HA)	2.1	1.7	1.9	2.0	2.7	1.1	1.7	1.1	1.8
Dwelling vacant/address not a dwelling (DV)	1.5	0.8	1.4	0.3	1.9	1.7	0.5	0.6	1.1
Dwelling destroy (DD)	0.1	0.1	0.0	0.3	0.0	0.1	0.3	0.0	0.1
Other (O)	0.3	0.2	0.1	0.0	0.5	0.4	0.2	0.3	0.3
Total	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Number of sampled households	1,998	3,416	736	931	1,171	840	1,036	700	5,414
Household response rate (HRR)[1]	99.2	99.5	99.4	99.2	99.1	99.1	99.9	99.6	99.4
Eligible men									
Completed (EMC)	92.6	92.6	94.8	91.7	88.9	93.8	93.8	93.8	92.6
Not at home (EMNH)	6.8	6.9	4.6	7.2	10.9	5.7	5.5	5.6	6.8
Refused (EMR)	0.5	0.0	0.0	0.6	0.1	0.3	0.1	0.0	0.2
Incapacitated (EMI)	0.1	0.4	0.6	0.0	0.1	0.2	0.5	0.6	0.3
Other (EMO)	0.0	0.2	0.0	0.5	0.0	0.0	0.1	0.0	0.1
Total	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Number of men	1,559	2,515	518	664	854	665	838	535	4,074
Eligible men response rate (EMRR)[2]	92.6	92.6	94.8	91.7	88.9	93.8	93.8	93.8	92.6
Overall response rate (OMRR)[3]	91.8	92.1	94.3	91.0	88.1	93.0	93.7	93.4	92.0

[1] Using the number of households falling into specific response categories, the household response rate (HRR) is calculated as:

$$\frac{100 * C}{C + HP + P + R + DNF}$$

[2] Using the number of eligible men falling into specific response categories, the eligible men response rate (EWRR) is calculated as:

$$\frac{100 * EMC}{EMC + EMNH + EMP + EWR + EMPC + EMI + EMO}$$

[3] The overall men response rate (OWRR) is calculated as:

$$OMRR = HRR * EMRR/100$$

5.5 Sampling errors

Sampling errors are important data quality parameters which give a measure of the precision of the survey estimates. The DHS survey final reports present sampling errors in Appendix B for selected indicators. The sampling error tables present the estimated indicator value, the standard error, the number of unweighted and weighted cases, the design effect, the relative standard error and the confidence limits. The design effect can be used in sample size calculation for subsequent survey designs. Section 4.2 deals with the details of the calculation of sampling errors; here we give an example of the national level sampling error table.

Table 5.9 Example table for sampling errors

Table B.2 Sampling errors for National sample, Bangladesh 2007

Variable	Value (R)	Standard error (SE)	Number of cases: Unweighted (N)	Number of cases: Weighted (WN)	Design effect (DEFT)	Relative error (SE/R)	Confidence limits: R-2SE	Confidence limits: R+2SE
			WOMEN					
Urban residence	0.226	0.006	10996	10996	1.539	0.027	0.213	0.238
No education	0.341	0.009	10996	10996	1.988	0.026	0.323	0.359
With secondary education or higher	0.363	0.009	10996	10996	1.972	0.025	0.344	0.381
Currently married	0.927	0.003	10996	10996	1.246	0.003	0.921	0.933
Currently pregnant	0.054	0.002	12951	13071	1.220	0.044	0.049	0.059
Children ever born	2.331	0.028	12951	13071	1.294	0.012	2.276	2.387
Children surviving	2.039	0.023	12951	13071	1.276	0.011	1.992	2.086
Children ever born to women age 40-49	4.566	0.061	2294	2254	1.349	0.013	4.444	4.689
Ever used any contraceptive method	0.830	0.007	10146	10192	1.841	0.008	0.817	0.844
Currently using any method	0.558	0.008	10146	10192	1.646	0.015	0.542	0.574
Currently using a modern method	0.475	0.008	10146	10192	1.633	0.017	0.458	0.491
Currently using pill	0.285	0.007	10146	10192	1.577	0.025	0.271	0.299
Currently using IUD	0.009	0.001	10146	10192	1.396	0.144	0.007	0.012
Currently using injectabs	0.070	0.004	10146	10192	1.670	0.060	0.062	0.079
Currently using female sterilization	0.050	0.004	10146	10192	1.681	0.073	0.043	0.057
Currently using periodic abstinence	0.049	0.003	10146	10192	1.262	0.055	0.044	0.054
Currently using withdrawal	0.029	0.002	10146	10192	1.222	0.071	0.025	0.033
Using public sector source	0.502	0.013	4751	4884	1.784	0.026	0.476	0.528
Want no more children	0.625	0.006	10146	10192	1.254	0.010	0.613	0.637
Want to delay at least 2 years	0.210	0.005	10146	10192	1.246	0.024	0.200	0.221
Ideal number of children	2.284	0.013	10756	10804	1.829	0.006	2.258	2.310
Mothers protected against tetanus for the last birth	0.902	0.008	4926	4905	1.843	0.009	0.886	0.918
Mothers received medical care at birth	0.180	0.009	6150	6058	1.685	0.050	0.162	0.198
Had diarrhea in the past 2 weeks	0.098	0.006	5789	5719	1.396	0.059	0.086	0.109
Treated with oral rehydration salts (ORS)	0.766	0.022	560	559	1.187	0.028	0.723	0.810
Sought medical treatment	0.198	0.021	560	559	1.194	0.104	0.157	0.239
Vaccination card seen	0.582	0.018	1161	1146	1.226	0.031	0.546	0.618
Received BCG vaccination	0.968	0.006	1161	1146	1.168	0.006	0.955	0.980
Received DPT vaccination (3 doses)	0.911	0.010	1161	1146	1.185	0.011	0.891	0.931
Received polio vaccination (3 doses)	0.908	0.010	1161	1146	1.213	0.011	0.888	0.929
Received measles vaccination	0.831	0.015	1161	1146	1.380	0.018	0.800	0.861
Received all vaccinations	0.819	0.016	1161	1146	1.365	0.019	0.787	0.850
Height-for-age (below -2SD)	0.432	0.010	5423	5312	1.379	0.023	0.412	0.451
Weight-for-height (below -2SD)	0.174	0.007	5423	5312	1.321	0.040	0.160	0.188
Weight-for-age (below -2SD)	0.410	0.009	5423	5312	1.291	0.022	0.392	0.428
BMI <18.5	0.297	0.007	9997	10021	1.569	0.024	0.282	0.311
Total fertility rate (past 3 years)	2.710	0.061	na	36507	1.313	0.022	2.589	2.832
Neonatal mortality (past 0-4 years)	36.677	3.296	6203	6103	1.218	0.090	30.085	43.270
Post-neonatal mortality (past 0-4 years)	14.826	2.025	6203	6094	1.254	0.137	10.775	18.877
Infant mortality (past 0-4 years)	51.503	3.942	6209	6108	1.259	0.077	43.620	59.386
Child mortality (past 0-4 years)	14.253	1.741	6255	6144	1.135	0.122	10.770	17.736
Under-five mortality (past 0-4 years)	65.022	4.387	6249	6144	1.264	0.067	56.248	73.796
Has heard of HIV/AIDS	0.674	0.011	10996	10996	2.557	0.017	0.651	0.697
Knows about condoms to prevent HIV/AIDS	0.319	0.008	10996	10996	1.893	0.026	0.302	0.336
Knows about limiting partners to prevent HIV/AIDS	0.325	0.009	10996	10996	2.045	0.028	0.307	0.343
			MEN					
Urban residence	0.227	0.007	3771	3771	1.040	0.031	0.213	0.242
No education	0.307	0.011	3771	3771	1.462	0.036	0.285	0.329
With secondary education or higher	0.365	0.012	3771	3771	1.495	0.032	0.342	0.389
Currently married	0.990	0.002	3771	3771	1.171	0.002	0.986	0.994
Ideal number of children	2.266	0.019	3624	3626	1.494	0.009	2.227	2.304
Has heard of HIV/AIDS	0.866	0.010	3231	3227	1.729	0.012	0.845	0.887
Knows condom use to prevent HIV/AIDS	0.658	0.012	3231	3227	1.453	0.018	0.634	0.683
Knows limiting partners to prevent HIV/AIDS	0.629	0.013	3231	3227	1.532	0.021	0.603	0.655

na = Not applicable

5.6 Sampling parameters in DHS data files

Some important sampling parameters should be included in the DHS final data set, such as domain, stratum, EA selection probability, and sampling weights. DHS survey final data files usually present geographic identifiers only down to domain or region level; district level identifiers are usually not presented due to confidentiality constraints. As for the sampling stratum identifier, DHS final data files should provide the true sampling stratum, which is important for many statistical analyses such as the sampling error calculation. However, in case of small strata having only a few clusters selected,

confidentiality constraints do not allow DHS data files to present the true sampling stratum identifier. In these cases, a higher level stratification identifier is included instead, which should be close to the true stratification and will not introduce substantial bias.

The standard sampling parameters included in the DHS Recode data files include:

1) Cluster indicator variable

2) Stratification variable

3) Sampling weight variables

4) Survey domain variables

5) First level geographical/administrative unit variable (region or province or department, etc.)

Glossary of terms

Analysis domain	A sub-population which cannot be identified in the sampling frame, such as domains specified by individual characteristics. See also *Design domain*.
Base map	A reference map that describes the geographic location and boundaries of an EA.
Cluster	The smallest geographic survey statistical unit for DHS surveys. It consists of a number of adjacent households in a geographic area. For DHS surveys, a cluster corresponds either to an EA or a segment of a large EA.
Collective living quarters	Living quarters such as army camps, boarding schools, or prisons where persons live individually. Collective living quarters are not considered as ordinary households and are excluded from DHS samples.
Confidence interval	A range within which the true value of an estimate likely lies. Usually reported as, with 95% confidence, the true value of $\bar{Y}$ will lie within the range of $\bar{y}-1.96*SE(\bar{y})$ and $\bar{y}+1.96*SE(\bar{y})$. Typically, DHS reports use $\bar{y}\pm2*SE(\bar{y})$ for a conservative estimate of 95% confidence limits.
Degrees of freedom	The number of independent units of information in a sample relevant to the estimation of a parameter or calculation of a statistic.
Design domain	A sub-population which can be identified in the sampling frame and therefore can be handled independently in the sample size and sampling procedures, usually consisting of geographic areas or administrative units. See also *Analysis domain*.
Design Effect (Deft)	A measure of efficiency of a complex sampling procedure compared to simple random sampling, defined as the ratio between the standard error using the given sample design and the standard error that would result if a simple random sample had been used.
Design weight	The inverse of the overall probability with which a sampling unit (household or individual) was selected in the sample. See also *Sampling weight*.
Desired precision	The level of accuracy of the results desired, often expressed as *Relative standard error* or coefficient of variation.
Dwelling unit	A room or a group of rooms normally intended as a residence for one household (for example: a single house, an apartment, a group of rooms in a house); a dwelling unit can have more than one household.

Enumeration Area (EA)	A geographic statistical unit which is created as a counting unit for a census and contains a certain number of households.
Explicit stratification	The actual division of the sampling units into specified parts known as strata. See also *Implicit stratification*.
Gross response rate	The number of households or individuals interviewed over the number selected.
Head of household	A person who is acknowledged as such by members of the household and who is usually responsible for the upkeep and maintenance of the household.
Household	A person or a group of related or unrelated persons, who live together in the same dwelling unit, who acknowledge one adult male or female 15 years old or older as the head of the household, who share the same housekeeping arrangements, and are considered as one unit.
Household listing	A complete listing of dwelling units/households in the selected EAs prepared prior to the selection of households.
Household selection	Random selection of the households from the household listing, typically by systematic selection.
Implicit stratification	The systematic sampling or probability proportional to size sampling of sampling units from an ordered list to achieve the effect of *Stratification*. See also *Explicit stratification*.
Item non-response	A sampling unit does not provide an answer for a specific question. See also *Unit non-response*.
Location map	A map produced in the household listing operation which indicates the main access to a cluster, including main roads and main landmarks in the cluster.
Master sample	A random sample of large size drawn from the census frame and prepared for use in a number of surveys, from which sub-samples can be selected for specific surveys.
Measure of size	A measurement reflecting the size of the sampling unit, typically the number of households or the total population of the sampling unit, available for each and every *primary sampling unit* in the country.
Non-sampling errors	Non-sampling errors result from problems during data collection and data processing, such as failure to locate and interview the correct household, misunderstanding of the questions on the part of either the interviewer or the respondent, and data entry errors.

Normalized standard weights	*Sampling weight* normalized by a constant factor such that the unweighted number of cases is the same as the weighted number of cases at the national level. Normalized standard weights are calculated for total households, total women and total men.
Primary Sampling Unit (PSU)	The *sampling unit* for the first stage of selection in a multi-stage sampling procedure; in DHS, typically an EA or a segment of an EA.
Probability sample	A sample in which the units are selected randomly with known and nonzero probabilities.
*Relative standard erro*r (RSE)	The amount of sampling error relative to the indicator level, independent of the scale of the indicator, calculated by dividing the standard error by the estimated value of the indicator
Sample take	The number of households or individuals to be interviewed per sample cluster.
Sampling errors	Sampling errors are the representative errors due to sampling of a small number of eligible units from the target population instead of including every eligible unit in the survey.
Sampling frame	A complete list of all sampling units that entirely covers the target population.
Sampling unit	The unit of selection at each stage of the sampling process. In a typical DHS with two-stage cluster sampling, the sampling unit at the first stage (*Primary sampling unit*) would be the EA, and the sampling unit at the second stage (*Secondary sampling unit*) would be the household.
Sampling weight	The design weight corrected for non-response or other calibrations.
Secondary Sampling Unit (SSU)	The sampling unit for the second stage of selection; in a typical DHS two-stage sample this is a household.
Self-weighting sample	A sample of individuals in which each individual has the same probability of being selected, and therefore a constant sampling weight is used. Also known as an *equal probability sample*.
Simple random sample (SRS)	A random selection of individuals or households drawn directly from the target population with each individual or household having equal probability of being selected.

Sketch map	A map produced in the household listing operation, with location of all structures found in the listing operation which helps the interviewer locate the selected households. A sketch map also contains the cluster identification information, location information, access information, and principal physical features and landmarks such as mountains, rivers, roads and electric poles.
SRSWOR	*Simple random sample* without replacement.
Standard error (SE)	The standard deviation of the sampling distribution of a statistic, or representative error due to sampling. See also *Sampling errors*.
Stratification	The process by which the survey population is divided into subgroups or strata that are as homogeneous as possible based on certain criteria. The principal objective of stratification is to reduce sampling errors.
Structure	A free-standing building or other construction that can have one or more units for residential or commercial use. Residential structures can have one or more dwelling units (for example: single house, apartment structure).
Student's t distribution	A family of continuous probability distributions that arises when estimating the mean of a normally distributed population in situations where the sample size is small and population standard deviation is unknown.
S*urvey domain/study domain*	A sub-population for which separate estimation of the main indicators is required.
Systematic selection (SYS)	Selection of units starting from a random point and selecting every n^{th} unit.
Target population	The population of interest in the survey, typically, in DHS, women age 15-49 and children under five years of age living in residential households. Most surveys also include men age 15-59.
Two-stage cluster sampling	At the first stage, a stratified sample of EAs is selected in each stratum, typically in DHS with probability proportional to size (PPS). At the second stage, a fixed (or variable) number of households is selected typically in DHS by equal probability systematic sampling.
Uniformly distributed random number	A random number which comes from a uniform distribution, that is, all possible values in the interval within which the random number is selected have equal probability of selection.
Unit non-response	A sampling unit (cluster, household, individual) is not interviewed at all. See also *Item non-response*.

Variance	A measure of how far a set of numbers is spread out around their mean.
Weight	An inflation factor which extrapolates the sample to the target population. See also *Design weight* and *Sampling weight*.

References

Aliaga, A. & Ren, R (2006). Optimal sample sizes for two-stage cluster sampling in Demographic and Health Surveys. DHS working papers No. 30.

Bankier, M. D. (1998). Power allocations: determining sample size for sub-national areas. *The American Statistician*, Vol. 42, PP. 174-177.

Cochran, W. G. (1977). *Sampling techniques*. John Wiley & Sons, New York

Deville, J.-C. & Särndal, C.-E. (1992). Calibration Estimators in Survey Sampling, *JASA*, Vol. 87, No. 418, pp. 376-382.

Deville, J.-C., Särndal, C.-E. & Sautory, O. (1993). Generalized Raking Procedures in Survey Sampling, *JASA*, Vol. 88, No. 423, pp. 1013-1020.

Dupont, F. (1994). Calibration Used as a Nonresponse Adjustment, IN: Diday, E. (ed.) *New Approaches in Classification and Data Analysis*, Springer Verlag, pp. 539-548.

Efron, B. & Tibshirani, R. J. (1993). *An introduction to the Bootstrap*. Chapman & Hall.

Hartley, H. O. & Rao, J. N. K. (1962). Sampling with unequal probabilities and without replacement. *Annals of Mathematical Statistics*, Vol. 33, pp. 350-374.

Kish, L. (1965). *Survey Sampling*. John Wiley & Sons, New York.

Lundström, S. & Särndal, C.-E. (1999). Calibration as a Standard Method for Treatment of non-response, JOS, Vol. 15, No. 2, pp.305-327.

Macro International Inc. (1996). *Sampling Manual*. DHS-III Basic Documentation No. 6. Calverton, Maryland.

Nieuwenbroek, N., Renssen R., Slootbeek, G. & Veugen, T. (1997). A General Weighting Package Including Estimates for Population Totals and Corresponding Variances: Extended Version, CBS Research Paper, No. 9745.

Platek, R. & Särndal, C.-E. (2001). Can a Statistician Deliver? *JOS*, Vol. 17, pp. 1-20.

Ren. R (2003). *Théories des sondages*. Lecture notes, ENSAI, France

Sautory, O. (1993). La macro SAS CALMAR: Redressement d'un Echantillon par Calage sur Marges, Document de travail de la Direction des Statistiques Démographiques et Sociales, no. F9310, INSEE.

Skinner, C. (1999). Calibration Weighting and Non-Sampling Errors, Research in Official Statistics, No. 1, pp.33-43.

Smith, T.M.F. (1990). Comment on Rao and Bellhouse: Foundations of survey based estimation and analysis. *Survey Methodology*, vol. 20, pp. 3-22.

Tillé, Y. (2001). *Théories des sondages.* Dunod, Paris.

Wolter, K. M. (1984). An investigation of some estimators of variance for systematic sampling. *JASA*, Vol. 79, pp. 781-790

Wolter, K. M. (1985). *Introduction to variance estimation*. Springer-Verlag, New York

Woodruff, R. S. (1971). A simple method for approximating the variance of a complicated estimate. *JASA*, Vol. 66, pp. 411-414.

Yates, F. & Grundy, P. M. (1953). Selection Without Replacement from Within Strata with Probability Proportional to Size. Journal of the Royal Statistical Society. Series B (Methodological) Vol. 15, No. 2 (1953), pp. 253-261. Blackwell Publishing.